Newton's method, zeroes of vector fields, and the Riemannian center of mass

David Groisser
Department of Mathematics
University of Florida
Gainesville FL 32611–8105
USA
groisser@math.ufl.edu

Revised version: July 2, 2003

Abstract

We present an iterative technique for finding zeroes of vector fields on Riemannian manifolds. As a special case we obtain a "nonlinear averaging algorithm" that computes the centroid of a mass distribution μ supported in a set of small enough diameter D in a Riemannian manifold M. We estimate the convergence rate of our general algorithm and the more special Riemannian averaging algorithm. The algorithm is also used to provide a constructive proof of Karcher's theorem on the existence and local uniqueness of the center of mass, under a somewhat stronger requirement than Karcher's on D. Another corollary of our results is a proof of convergence, for a fairly large open set of initial conditions, of the "GPA algorithm" used in statistics to average points in a shape-space, and a quantitative explanation of why the GPA algorithm converges rapidly in practice; see [11].

We also show that a mass distribution in M with support Q has a unique center of mass in a (suitably defined) convex hull of Q.

2000 AMS Subject Classification: Primary 53B21, 60D05; secondary 53C99

Key Words: nonlinear averaging, center of mass, centroid, convex hull, Procrustean mean, shape space

1 Introduction.

In this article we present an iterative technique for finding zeroes of vector fields on Riemannian manifolds, and apply this technique to the averaging of a mass distribution with support contained in a sufficiently small ball in a Riemannian manifold. Our approach provides a new and constructive proof of Karcher's theorem on the existence and uniqueness of the center of mass, under a somewhat stronger requirement on the radius of the supporting ball than was used in [16].

This study was originally motivated by curiosity about a method (the "GPA algorithm") used in statistics to find the average, suitably defined, of a sample of shapes. In many areas of image analysis, particularly in biological applications such as cardiography (cf. [27]) and maps of the brain (cf. [1]) this average is the starting point for understanding "normal" shapes and deviations from the norm. In practical applications the averaging algorithm tends to converge remarkably quickly, often stabilizing to desired precision after two or three iterations (cf. [1], Figure 5 (p. 22), or [8] Table 3 (p. 307)). The initial purpose of our study was to understand the geometry underlying this algorithm and, in quantitative terms, why the convergence in practical applications is so rapid. In exploring this the author found that the GPA algorithm has a more general interpretation on Riemannian manifolds, generalizing to a technique for finding local zeroes of a vector field. The technique is an iterative algorithm that we show is closely related to Newton's method and mimics the contracting-mapping proof of the Inverse Function Theorem.

As a special case of this technique, we obtain a general Riemannian averaging algorithm. The vector field used in this algorithm has a unique local zero, assuming the diameter D of the support of distribution being averaged is not too large, and is "almost linear" near this zero if D is small, explaining the rapid convergence. This zero is exactly the Riemannian center of mass of the distribution being averaged. In sections 4 and 5 of this paper we quantify "not too large" and "small", giving sufficient conditions for convergence of the algorithm and estimating the convergence rate.

The Riemannian averaging algorithm can in principle be applied to any "nonlinear averaging" problem in which the objects being averaged are parametrized by a Riemannian manifold, and is easily implemented in spaces for which the exponential map and its inverse are explicitly known (e.g. Riemannian submersions from spheres, and certain homogeneous spaces with invariant metrics). This is exactly the situation for the shape-averaging problem. The (Euclidean) *shape space* Σ_n^k is the space of configurations of k non-identical labeled points in $\mathbf{R}^n$, modulo equivalence under translations, rotations, and dilations (rescalings) in $\mathbf{R}^n$; sometimes one also allows reflections. The *size-and-shape space* $\tilde{\Sigma}_n^k$ is defined similarly, but one does not mod out by rescalings. These spaces can naturally be given the structure of manifolds with singularities, with natural Riemannian metrics on their smooth parts ([17, 2, 18]). Averaging (sizes-and-) shapes can be viewed as averaging certain mass distributions

on (size-and-) shape spaces, namely finite lists of points with normalized counting measure. In the probability and statistics literature there is a commonly accepted definition of mean size-and-shape, the Procrustean mean size-and-shape, but several possible definitions of mean shape (see [21] p. 292 and [22]), the most common of which may be the Procrustean mean shape used in [23]. However, while the Procrustean mean *size-and-shape* as defined in the probability and statistics literature agrees with the Riemannian center of mass, the Procrustean mean *shape* does not.

The GPA (Generalized Procrustes Analysis) algorithm as described in [23] lives intrinsically on size-and-shape space; call this algorithm GPA-SS. To obtain from this an algorithm that averages shapes, one first embeds shape-space Σ_n^k into size-and-shape space $\tilde{\Sigma}_n^k$ in a standard way, carrying the list Q of shapes to be averaged to a list $\iota(Q)$ of sizes-and-shapes. One then produces a sequence of in $\tilde{\Sigma}_n^k$ by applying the GPA-SS algorithm to $\iota(Q)$. Finally one projects the limit (if there is one) back onto shape space. Call this set of steps GPA-S. Le proves in [23] that if the shapes in Q are not too far apart in Σ_n^k, and if the sequence in $\tilde{\Sigma}_n^k$ converges, then the limit in $\tilde{\Sigma}_n^k$ is the Procrustean mean size-and-shape of the list $\iota(Q)$. It is not hard to show that this projection of the Procrustean mean size-and-shape is exactly the Procrustean mean shape ([23], p. 54), so that GPA-S computes the Procrustean mean shape.

Although the literature contains many discussions of the GPA-SS and other GPA-derived algorithms, at the time this paper was first completed [10] the literature contained no theorems giving sufficient conditions for any of these algorithms to converge. However, as we show in [11], the GPA-SS algorithm is *exactly* our Riemannian averaging algorithm as applied to size-and-shape space. Hence convergence of the GPA-SS algorithm, for an explicitly describable open set of initial conditions, is an immediate corollary of the Riemannian-averaging theorems in sections 4 and 5 of this paper. After [10] was written, [25], which contains some overlapping results, appeared.

In the iterative part of the GPA-S algorithm, one can obtain a sequence of points in shape space by projecting each point in the GPA-SS sequence, rather than just the limit, back onto shape space. (This sequence in shape space can also be described slightly more intrinsically; see [11], where we discuss the application of the results of this paper to Procrustean averaging in more detail.) In this way one obtains an iterative algorithm GPA-S$'$ on shape space itself. GPA-S$'$ does not coincide with the Riemannian averaging algorithm on shape space—it cannot, since it converges (for suitable initial conditions) to the Procrustean mean shape and not to the Riemannian average. However, GPA-S$'$ is an algorithm of the more general type also considered here, and therefore its convergence, again for an explicitly describable open set of initial conditions, follows directly from our more general theorems in section 2, as well as from the fact that GPA-S converges.

In this paper we also address the question of why the convergence of the GPA algorithms is so rapid in practice. As has been noted by many authors, the data sets averaged in practical applications tend to be very concentrated sets in shape (or size-and-shape) space; their diameter D is very small compared with any length-scale

derivable from the geometry of shape (or size-and-shape) space. Our theorems in section 5 show why, for small D, convergence is rapid.

To describe our results more concretely, we need some notation and terminology:

Definition 1.1 Let $(A, d_A), (B, d_B)$ be metric spaces and let $\kappa \in [0, 1)$. We call a map $F : A \to B$ a *contraction with constant* κ if $d_B(F(x), F(y)) \leq \kappa\ d_A(x, y)$ for all $x, y \in A$.

The results of this paper are proved using a version of the Contracting Mapping Theorem (Theorem 2.1). The maps we use arise from certain vector fields, perhaps defined only locally, on Riemannian manifolds. To describe these maps, let ∇ be the Levi-Civita connection on a Riemannian manifold (M, g), not assumed complete. If X is a C^1 vector field defined on some open set $V \subset M$, then at each point $p \in V$ we can view the covariant derivative ∇X as a linear transformation $T_pM \to T_pM$, namely $v \mapsto \nabla_v X$. Call X *nondegenerate* on a subset $U \subset V$ if this endomorphism $(\nabla X)_p$ is invertible for all $p \in U$. When referring to bounds on $(\nabla X)_p^{-1}$ and other linear transformations, throughout this paper we use the operator norm: $\|T\| = \sup_{\|v\|=1} \|T(v)\|$.

A C^1 vector field X defined on an open set in M and nondegenerate on a subset U defines a map $\Phi_X : U \to M$ by

$$\Phi_X(p) = \exp_p(-(\nabla X)_p^{-1} X_p), \tag{1.1}$$

assuming that $\exp_p(-(\nabla X)_p^{-1} X_p)$ is defined for all $p \in U$. (In this paper we use both X_p and $X(p)$ to denote the value of a vector field X at at a point p.) Note that zeroes of X are fixed-points of Φ_X, and if $\|X\|$ is not too large pointwise then the converse is true as well. One of the results of this paper is the following theorem, a much stronger version of which is proven in §2.

Theorem 1.2 *Let (M, g) be a Riemannian manifold and let $U \subset M$ be open. Given $\epsilon > 0, k_1 > 0, k_2 > 0$, let $\mathcal{X}_{\epsilon,k_1,k_2}(U)$ denote the set of nondegenerate vector fields X on U satisfying the following conditions pointwise on U: (i) $\|X\| \leq \epsilon$, (ii) $\|(\nabla X)^{-1}\| \leq k_1^{-1}$, and (iii) $\|\nabla\nabla X\| \leq k_2$. If both ϵk_1^{-1} and $k_2 k_1^{-1}$ are sufficiently small, and $X \in \mathcal{X}_{\epsilon,k_1,k_2}(U)$, then $\Phi_X : U \to M$ is a contraction, where the distance function on U is the one determined by the Riemannian metric g on M. If U is a ball B of radius ρ centered at p_0, and if ρ is sufficiently small and ϵ, k_1, k_2 are as above, then there exists a positive $\epsilon_1 \leq \epsilon$ such that if $\|X(p_0)\| \leq \epsilon_1$, then Φ_X preserves B and hence has a unique fixed point $\overline{p}$ in B; the point $\overline{p}$ is also the unique zero of X in B. For all p in some possibly smaller open ball centered at p_0, the iterates $(\Phi_X)^n(p)$ converge to $\overline{p}$.*

Example 1.3 *Euclidean space* $\mathbf{R}^n$. Since $T_x\mathbf{R}^n \cong \mathbf{R}^n$ canonically for all $x \in \mathbf{R}^n$, a vector field X on $\mathbf{R}^n$ can be naturally identified with a vector-valued function $G : \mathbf{R}^n \to \mathbf{R}^n$, and the Levi-Civita connection is just given by ordinary directional

differentiation: $(\nabla X)_x(v) = (DG|_x)(v) = \frac{d}{dt}G(x+tv)|_{t=0}$. The exponential map is given simply by $\exp_x(v) = x + v$. Thus

$$\Phi_X(x) = x - (DG|_x)^{-1}(G(x)),$$

which is exactly the Newton's-method map used in the usual contracting-mapping proof of the Inverse Function Theorem; cf. [26] §4.9.

Example 1 illustrates the close relationship between the iteration in Theorem 1.2 and Newton's method. However, one gains considerable flexibility by not requiring quite so strict a relationship as in (1.1), looking more generally at maps of the form $p \mapsto \exp_p(Y_p) := \Psi_Y(p)$ for suitable vector fields Y. Our approach will focus on maps of this more general form, deducing consequences for maps of the form Φ_X as a special case. For the maps Ψ_Y, the size restriction on $\|\nabla X\|$ and $\|\nabla\nabla X\|$ is replaced by the single condition that at each point the endomorphism ∇Y be close to minus the identity. Note that in this case, $-(\nabla Y)^{-1}Y$ is close to Y, so that the maps Φ_Y and Ψ_Y are themselves close. Iterative schemes based on maps of the form Ψ_Y are thus a natural generalization of Newton's method. Our most general result for these maps and their associated algorithms is Theorem 2.8, a stronger version of Theorem 1.2 in which all the "sufficiently smalls" are quantified for the maps Ψ_Y and Φ_X. One corollary is the following:

Corollary 1.4 *Let $\delta \le \Delta \in \mathbf{R}, r_1 \in \mathbf{R}$, and suppose that the sectional curvature K of M satisfies $\delta \le K \le \Delta$. There exists a number D_{crit}, depending only on δ, Δ, and r_1, such if μ is a probability distribution supported on a set $Q \subset M$ of diameter less than D_{crit}, and the local convexity radius at some point of Q is at least r_1, then the primary center of mass $\overline{q}$ of μ exists, and the Riemannian averaging algorithm converges to $\overline{q}$ for every initial point $q \in Q$.*

The definition of D_{crit} in terms of δ, Δ, and r_1 is given in §4 (see (4.18)); the "primary center of mass" is defined in §3.

We use the exponential map in defining Ψ_Y because of its universality, but in specific examples "exp" can be replaced by other maps defined on a neighborhood of the zero-section of the tangent bundle. This is convenient in the shape-space setting for the algorithm GPA-S′; see [11]. However, any continuous map $F : (U \subset M) \to M$ can always be expressed in the form $\exp \circ Y$, with Y continuous, provided that for all $p \in U$ the distance $d(p, F(p))$ is less than the local injectivity radius at p (see Definition 2.4). Thus if we are interested only in maps that have any chance of having fixed points, we can always restrict attention to maps of the form Ψ_Y.

This paper is organized as follows. In §2 we study the maps Ψ_Y and derive conditions for iterative algorithms based on these maps to converge. Before specializing

to the Riemannian averaging algorithm, some discussion of Riemannian centers of mass is needed; this is given in §3, where we also define the vector field Y on which the averaging algorithm is based. In general a probability distribution on a manifold (even one supported on a finite set) can have more than one center of mass, depending on how "center of mass" is defined, but under certain circumstances one of these is distinguished. In statistics this is typically done using least-squared-distances minimization. However, we offer a more directly geometric way of singling out a "primary" center of mass, using convex hulls. We digress a bit in Section 3 from the main contracting-mapping theme because, surprisingly, we have not found any discussion of the relation of Riemannian centers of mass to convex hulls anywhere in the center-of-mass literature, although the idea seems very natural. Our final statement concerning this relationship, Corollary 3.13, may be a fact known to workers in the field but it is a stronger explicit statement than we have seen elsewhere.

In §4 we apply the results of §2 to obtain a constructive proof of the existence and uniqueness of the center of mass of a probability distribution μ with sufficiently support in a ball of sufficiently small radius ρ (Corollary 4.7). Karcher's existence/uniqueness theorem has a less stringent requirement on ρ, and its uniqueness statement has been strengthened by W. S. Kendall [19]. In view of these results, the most important feature of the contracting-mapping approach to the center-of-mass problem is not that it gives existence and uniqueness of the average, but that it provides a constructive algorithm for finding it (Theorem 4.8), along with convergence-rate estimates. The restriction on ρ in Theorem 4.8 is almost certainly not sharp. If the map on which the algorithm is based has a certain convexity property that we call "tethering", then the upper limit on ρ can be increased considerably. Tethering may occur fairly generally, but the author has no proof of this. Thus the results in sections 4–6 are stated both without and with the assumption of tethering.

In §5 we estimate the convergence rate of algorithms of the form "iterate Ψ_Y" for general Y, and show that the rate is completely controlled by bounds on $\nabla Y + I$. In general the convergence of the sequence $\{p_n = \Psi_Y(p_0)\}$ is exponential; if $\|\nabla Y + I\| \leq \epsilon_1$ then $d(p_{n+1}, p_n) \leq d(p_1, p_0)\epsilon_1^n$. For maps of the form Φ_X the convergence is much faster, obeying the same bounds that one has for Newton's method in Euclidean space. For the Riemannian averaging algorithm we obtain something in between: exponential convergence, but with a constant ϵ_1 that is $O(D^2)$, where D is the diameter of the support of the distribution being averaged. We also combine the convergence-rate result with W. S. Kendall's uniqueness result to obtain a sharpening of Theorem 4.8 (Theorem 5.3), establishing convergence of the algorithm under a weaker requirement on ρ.

The statement that ϵ_1 is $O(D^2)$ heuristically—and *only* heuristically—explains the rapid convergence of the GPA algorithms; it does not fully explain why GPA algorithms converges rapidly in any applications (or determine in advance whether they will), since asymptotics do not tell us how small D must be before the leading asymptotic term decently approximates the actual convergence rate. However, Theo-

rem 5.3 can be used to give bounds on ϵ_1 of the form $\epsilon_1 \leq cD^2$ (for all D less than the critical diameter in the theorem, not just for small D), where c is computable from the geometry of M. In §6 we carry this out and give a *universal* worst-case estimate of the convergence rate when the curvature of M is non-negative, which is the case in all shape space and size-and-shape space applications.

In the appendix (§7) we prove (or cite proofs of) certain facts used in §§2–4 concerning Jacobi fields and the distance function.

2 Zeroes of Vector Fields

Throughout this paper, M denotes a smooth connected manifold equipped with a Riemannian metric g. The induced distance function on $M \times M$ is denoted $d_M(\cdot, \cdot)$, or simply $d(\cdot, \cdot)$ when no ambiguity can arise. M is always regarded as a metric space with this distance function, and the closure of a subset U in M is denoted $\overline{U}$. $B_\rho(p) \subset M$ denotes the open ball of radius ρ centered at p. If $U \subset M$ is connected, d_U denotes "distance within U", the infimum of lengths of curves in U connecting two given points of U. TM denotes the tangent bundle of M, and $\pi : TM \to M$ the canonical projection. X and Y denote vector fields on M that are at least C^2 and C^1 respectively. If N_1, N_2 are manifolds and $F : N_1 \to N_2$ is a smooth map, then for $p \in N_1$, we let $F_{*p} : T_pN_1 \to T_{F(p)}N_2$ denote the derivative of F at p. The identity map of any space is denoted I.

The main theorems of this paper are deduced from the following corollary of the standard Contracting Mapping Theorem (cf. [26] Corollary 4.9.2).

Theorem 2.1 (Contracting Mapping Theorem) *Let $B = B_\rho(p_0)$ be an open ball in a metric space (A, d), with $(\overline{B}, d)$ complete. Suppose that $B \subset U \subset A$, that $F : U \to A$ is a contraction with constant κ, and that $d(p_0, F(p_0)) < (1-\kappa)\rho$. Then F preserves B and has a unique fixed point $\overline{p}$. Furthermore $\overline{p} \in B$ and $\lim_{n\to\infty} F^n(q) = \overline{p}$ for all $q \in B$.*

As in the Euclidean case (Example 1), in the general case Φ_X (and more generally Ψ_Y) turns out to be a contraction on sets on which $\|X\|$ (more generally $\|Y\|$) is sufficiently small. Our proof of this fact relies on the following simple fact.

Lemma 2.2 *Let U, M be connected Riemannian manifolds and let $\kappa < 1$. If $F : U \to M$ is a C^1 map satisfying*

$$\|F_{*p}\| \leq \kappa \quad \text{for all } p \in M \tag{2.1}$$

then F is a contraction with constant κ.

Proof: For any curve γ in U connecting p to q, (2.1) implies $\ell(F \circ \gamma) \leq \kappa\ell(\gamma)$, where ℓ denotes arclength. ∎

We will prove that Φ_X is a contraction (on suitable sets) by computing its derivative and applying Lemma 2.2. The map Φ_X is of the form $\exp \circ Y$, where Y is a vector field on M. Below we express the derivatives of the maps $Y : M \to TM$ and $\exp : TM \to M$ in terms of the horizontal-vertical splitting of $T(TM)$ induced by the Levi-Civita connection ∇. We first review this splitting (see also [16], Appendix B).

Given a curve γ in M starting at a point p (i.e. a map γ from some interval of the form $(-\epsilon, \epsilon)$ to M with $\gamma(0) = p$), a *lift* of γ starting at $w \in T_pM$ is a curve $\tilde{\gamma}$ with $\pi \circ \tilde{\gamma} = \gamma$ and $\tilde{\gamma}(0) = w$—i.e. a vector field along γ whose value at p is w. A lift $\tilde{\gamma}$ is *horizontal* if this vector field is parallel ($\nabla_{\gamma'(t)}\tilde{\gamma} \equiv 0$). Every curve γ has a unique horizontal lift starting a a given $w \in T_{\gamma(0)}M$, and the vector $\tilde{\gamma}'(0) \in T_w(TM)$ depends only on $\gamma'(0)$. Hence the map $\gamma'(0) \mapsto \tilde{\gamma}'(0)$ is well-defined and at each $w \in TM$ uniquely determines a *horizontal lift* $\tilde{v} \in T_w(TM)$ of each $v \in T_pM$, where $p = \pi(w)$. The *horizontal subspace* of $T_w(TM)$ is defined to be the subspace H_w consisting of all horizontal lifts to w of vectors in T_pM, and $\pi_{*w}|_{H_w} : H_w \to T_pM$ is an isomorphism. The *vertical subspace* V_w of $T_w(TM)$ is the tangent space to the fiber T_pM at w. The subspace V_w is canonically isomorphic to T_pM (identifying a vertical vector $\frac{d}{dt}(w + tv)|_{t=0}$ with v); we denote the inverse of this isomorphism by $\iota : T_pM \to V_w(TM)$. The horizontal and vertical subspaces provide a splitting of $T_w(TM)$: for every $u \in T_w(TM)$, there exist unique vectors $a, b \in T_pM$ such that $u = \tilde{a} + \iota(b)$ (specifically $a = \pi_* w$ and $b = \iota^{-1}(w - \tilde{a})$); we write $\tilde{a} = \mathrm{hor}(u)$ and $\iota(b) = \mathrm{vert}(u)$.

The derivatives we need will be expressed in terms of Jacobi fields (vector fields J along geodesics γ satisfying the Jacobi equation

$$\nabla_{\gamma'}\nabla_{\gamma'}J = \mathrm{Riem}(\gamma', J)\gamma'; \tag{2.2}$$

see [4] §1.4 or [16] Appendix A). Below, for $w, a, b \in TM$ with the same base-point, let $J^w_{(a,b)}$ denote the Jacobi field J along γ_w with $J(0) = a, (\nabla_{\gamma'_w}J)(0) = b$, where γ_w is the unique geodesic with initial velocity w (i.e. $\gamma_w(t) = \exp_{\pi(w)}(tw)$).

Throughout this paper we will be concerned with maps of the form

$$\Psi_Y = \exp \circ Y : U \to M, \tag{2.3}$$

where Y is a vector field on some domain $U \subset M$. In (2.3) we view Y as a map $U \to TM$ and assume that image$(Y) \subset$ domain(exp). The derivative $(\Psi_Y)_*$ is given by

$$(\Psi_Y)_{*p}v = J^w_{(v,0)}(1) + (\exp_p)_{*w}(\iota((\nabla_v Y)_p)) \ \in T_{\Psi_Y(p)}M \tag{2.4}$$

where $w = Y_p$; the formula above can be deduced from [16] Appendix B. If X is a nondegenerate vector field and we define $\Phi_X : U \to M$ as in (1.1), then for the vector field $Y = -(\nabla X)^{-1}X$ we have

$$\nabla Y = (\nabla X)^{-1} \circ (\nabla\nabla X) \circ (\nabla X)^{-1}X - I. \tag{2.5}$$

Thus as a particular case of (2.4) we have

$$(\Phi_X)_{*p}(v) = J^{Y_p}_{(v,0)}(1) - (\exp_p)_{*Y_p}(\iota(v)) + (\exp_p)_{*Y_p}(\iota((\nabla X)^{-1} \circ (\nabla_v \nabla X) \circ (\nabla X)^{-1} X)), \tag{2.6}$$

where $X, \nabla X$, and $\nabla_v \nabla X$ are evaluated at p.

The term $(\exp_p)_{*Y_p}(\iota(v))$ in (2.6) is itself the value of a Jacobi field, namely $J^w_{(0,v)}(1)$ where $w = Y_p$. Hence (2.6) can be written as

$$(\Phi_X)_{*p}(v) = \hat{J}^p_v(1) + (\exp_p)_{*Y_p}(\iota(Z_p)) \tag{2.7}$$

where $Z_p = (\nabla X)^{-1}_p (\nabla_v \nabla X)_p (\nabla X)^{-1}_p X_p$ and where $\hat{J}^p_v$ is the Jacobi field along γ_w with the "antidiagonal" initial conditions $\hat{J}(0) = -(\nabla_{\gamma'} \hat{J})(0) = v$. In Euclidean space this Jacobi field always vanishes at time 1, and $(\exp_p)_*$ is the identity after appropriate identifications are made as in Example 1, so that (as is well known) Φ_X is a contraction if at each point $\|X\|$ is small enough in terms of $\|(\nabla X)^{-1}\|$ and $\|\nabla \nabla X\|$. In the general case we can again make $\|(\exp_p)_*(\iota(Z_p))\|$ arbitrarily small by taking $\|X_p\|$ sufficiently small. Additionally, $\|X_p\|$ small implies $\|Y_p\|$ small, implying that the geodesic γ_{Y_p} is short. For sufficiently short geodesics, the map $v \mapsto \|\hat{J}^p_v(1)\|$ is arbitrarily close to the corresponding map on Euclidean space, namely the zero map. (We will prove a stronger version of this fact in Lemma 2.3 below.) Hence it is already clear that if $\sup_p \|X_p\|$ is sufficiently small on a set U, then $(\Phi_X)|_U$ will be a contraction.

The essential ingredient in the preceding argument is that Φ_X is a map of the form $\Psi_Y = \exp \circ Y$ for some vector field Y whose covariant derivative is close to minus the identity (pointwise) whenever $\|Y\|$ is small enough. (The prototypical example is the radial vector field $-\sum_i x^i \frac{\partial}{\partial x^i}$ on $\mathbf{R}^n$, whose covariant derivative is identically $-I$.) In computational situations it may be costly to invert ∇X, so we will analyze the more general maps Ψ_Y, and deduce results for maps of the form Φ_X as a special case.

For some applications (e.g. those in [11]), it is useful to know the explicit dependence of our eventual contraction constants on background geometric parameters, so we keep track of this dependence carefully—leading unavoidably to longer formulas than if we were only aiming at qualitative results. Certain special functions will appear, all of which are related to the analytic (entire) functions $\mathbf{c}, \mathbf{s}$ defined by

$$\mathbf{c}(z) = \sum_{n=0}^{\infty} \frac{z^n}{(2n)!}, \quad \mathbf{s}(z) = \sum_{n=0}^{\infty} \frac{z^n}{(2n+1)!}. \tag{2.8}$$

Since the definitions and properties of the relevant functions are scattered through the text, for reference Table 1 lists the functions and the properties used.

To estimate $\|(\Psi_Y)_*\|$, we rewrite (2.4) as

$$(\Psi_Y)_{*p} v = \hat{J}^p_v(1) + (\exp_p)_{*Y_p}(\iota((\nabla Y|_p + I)v)). \tag{2.9}$$

Table of Special Functions

function	defining formula	properties used
$\mathbf{c}(z),\ z \in \mathbf{C}$	$\sum_{n=0}^{\infty} z^n/(2n)!$	
$\mathbf{s}(z),\ z \in \mathbf{C}$	$\sum_{n=0}^{\infty} z^n/(2n+1)!$	
$\phi_-(x),\ x \in [0,\infty)$	$\mathbf{c}(x^2) - \mathbf{s}(x^2)$ $= \cosh x - x^{-1}\sinh x$	mono. $\uparrow$, $\phi_-(x) \geq 0$
$\phi_+(x),\ x \in [0, 3\pi/4)$	$\mathbf{s}(-x^2) - \mathbf{c}(-x^2)$ $= x^{-1}\sin x - \cos x$	mono. $\uparrow$, $\phi_+(x) \geq 0$
$C_1(\lambda, r), \lambda \in \mathbf{R}, r \geq 0$	1, if $\lambda \geq 0$ $\frac{\sinh(\lvert\lambda\rvert^{1/2}r)}{\lvert\lambda\rvert^{1/2}r}$, if $\lambda < 0$	mono. $\uparrow$ in each variable, $C_1(\lambda, r) \geq 1$
$h(\lambda, r),\ \lambda \in \mathbf{R}$, $r \in \begin{cases} [0,\pi) \text{ if } \lambda > 0, \\ [0,\infty) \text{ if } \lambda \leq 0 \end{cases}$	$\mathbf{c}(-\lambda r^2)/\mathbf{s}(-\lambda r^2)$	$h(0,r) = h(\lambda, 0) = 1$, $h(\lambda, r) > 0$ if $\lambda \leq 0$, or if $\lambda > 0$ and $\lambda^{1/2} r < \pi/2$
$h_-(x) = h(-1, x),\ x \in [0,\infty)$	$x \coth x$	mono. $\uparrow$, $h_-(x) \geq h_-(0) = 1$
$h_0(x) = h(0, x),\ x \in [0,\infty)$	1	
$h_+(x) = h(1, x),\ x \in [0,\pi)$	$x \cot x$	mono. $\downarrow$, $h_+(x) \leq h_+(0) = 1$
$\psi(\lambda, r)$, same domain as h	$\operatorname{sign}(\lambda)(1 - h(\lambda, r))$	$\psi(\lambda, r) \geq 0$, mono. $\uparrow$ in $\lvert\lambda\rvert$ and r, convex in each variable
$\psi_{\max}(\delta, \Delta, r),\ \delta \leq \Delta \in \mathbf{R}$, $r \in [0,\infty)$	$\max(\psi(\Delta, r), \psi(\delta, r))$	mono. $\uparrow$ in Δ and r, mono. $\downarrow$ in δ, convex in each variable, $\psi_{\max}(\delta, \Delta, 0) = 0$

Table 1: In this table and throughout this paper our convention for functions that are given for $x \neq 0$ by formulas such as "$x^{-1}\sin x$)" are extended to $x = 0$ by continuity. When monotonicity or convexity of a multivariable function is stated with respect to one variable, the other variables are assumed fixed.

We will first analyze the Jacobi fields $\hat{J}_v^p$.

Notation. For any subset $U \subset M$, let $\Delta(U)$ and $\delta(U)$ denote, respectively, the supremum and the infimum of the sectional curvatures of $(U, g|_U)$; let $|K|(U) = \max(|\Delta(U)|, |\delta(U)|)$. For a curve γ we simply write $\Delta(\gamma)$ for $\Delta(\mathrm{Im}(\gamma))$, etc. Then we have the following proposition. The inequality (2.10) below can be derived from Karcher's elegant (and more general) Jacobi-field bounds; see [16] pp. 534-535, 539. However, for the special case (2.10), we give a short, direct proof in the Appendix (§7.1). In the second part of §7.1 we show how the proof leads directly to (2.12).

Proposition 2.3 *Let $p \in M$, let $\gamma : [0,1] \to M$ be a geodesic of length r starting at p, and for each $v \in T_pM$ let $\hat{J}_v$ be the Jacobi field along γ with the "antidiagonal" initial conditions $(\hat{J}_v(0), (\nabla_{\gamma'}\hat{J}_v)(0)) = (v, -v)$. Let $v^\perp$ denote the component of v perpendicular to $\gamma'(0)$. Then*

$$\|\hat{J}_v(1)\| \leq \phi_-(r|K|(\gamma)^{1/2})\|v^\perp\| \tag{2.10}$$

where

$$\phi_-(x) = \cosh(x) - \frac{\sinh x}{x}. \tag{2.11}$$

If M is a locally symmetric space of nonnegative curvature, and $\Delta(\gamma)^{1/2}r < 3\pi/4$, this bound can be sharpened to

$$\|\hat{J}_v(1)\| \leq \phi_+(r\Delta(\gamma)^{1/2})\|v^\perp\| \tag{2.12}$$

where

$$\phi_+(x) = \frac{\sin x}{x} - \cos x. \tag{2.13}$$

The "$3\pi/4$" in the locally-symmetric case can be increased to approximately $.87\pi$ (see the discussion of (7.5) in §7.1)) but any instances in which $r\Delta(\gamma)^{1/2} > \pi/2$ are irrelevant for all uses in this paper.

Turning our attention to the second term in (2.9), we have

$$\|(\exp_p)_{*Y_p}(\iota(\nabla Y|_p + I)v)\| \leq \|(\exp_p)_{*Y_p}\| \, \|(\nabla Y|_p + I)\| \, \|v\|.$$

We recall the following terminology.

Definition 2.4 The *local injectivity radius* at $p \in M$ is $r_{\mathrm{inj}}(p) := \sup\{\rho \mid \exp_p : (B_\rho(0) \subset T_pM) \to M$ is defined and is a diffeomorphism onto its image$\}$; $r_{\mathrm{inj}}(\cdot)$ is a positive continuous function on M. For any subset $U \subset M$, we define $r_{\mathrm{inj}}(U) = \inf_{p\in U}\{r_{\mathrm{inj}}(p)\}$. When $U = M$ this infimum is called the *injectivity radius of* (M, g).

Definition 2.5 A subset $U \subset M$ is *convex* (respectively, *strongly convex*) if for all $p, q \in U$ (resp., for all $p \in U, q \in \overline{U}$) there is a unique minimal geodesic segment γ in M from p to q, and $\gamma - \{q\}$ lies entirely in U[1]. For each $p \in M$ we define the *local convexity radius* $r_{\text{cvx}}(p) := \sup\{\rho \leq r_{\text{inj}}(p) \mid B_\rho(p) \text{ is convex}\}$; for $U \subset M$ we let $r_{\text{cvx}}(U) = \inf_{p \in U}\{r_{\text{cvx}}(p)\}$. Like the local injectivity radius, the local convexity radius of a point (or of a closed set) is always positive ([13] Lemma I.6.4).

Convexity is relevant because we want to apply Lemma 2.2 and Theorem 2.1 to the case $U \subset M$. The lemma only gives us a contraction from the metric space (U, d_U) to (M, d_M). However if U is convex then $d_U = d_M$ so that Theorem 2.1 applies.

For $w \in T_pM$ with $\|w\| < r_{\text{inj}}(p)$ the norm of $(\exp_p)_{*w}$ can be bounded in terms of curvature and $\|w\|$:

$$\| \exp_{p*w} \| \leq C_1(\delta(\gamma), \|w\|) \tag{2.14}$$

where γ is the geodesic from p with $\gamma'(0) = 1$ and where

$$C_1(\lambda, r) = \begin{cases} 1, & \lambda \geq 0 \\ \frac{\sinh(|\lambda|^{1/2} r)}{|\lambda|^{1/2} r}, & \lambda < 0 \end{cases} \tag{2.15}$$

(see [16] estimate C1). Thus if the image of γ lies in a set U, and $\|Y_p\| < r_{\text{inj}}(p)$, then

$$\|(\exp_p)_{*Y_p}(\iota(\nabla Y|_p + I)v)\| \leq C_1(\delta(U), \|Y_p\|)\ \|(\nabla Y + I)_p\|\ \|v\|. \tag{2.16}$$

Assembling the pieces above, we have the following corollary.

Corollary 2.6 *Let (M, g) be a Riemannian manifold, $\rho > 0$, $p \in M$, and $B = B_\rho(p)$. Assume that $\rho \leq r_{\text{cvx}}(B)$ and that $|K|(B) < \infty$.*

(a) There exists $\epsilon > 0$, depending only on $r_{\text{cvx}}(B)$ and the sectional curvature of (B, g), such that if Y is a vector field defined on B, with $\|Y\| \leq \epsilon$ and $\|\nabla Y + I\| \leq \epsilon$ pointwise on B, then Y has a unique zero in B, namely $\lim_{n\to\infty}(\Psi_Y)^n(q)$ for any $q \in B$.

(b) Let $k_1, k_2 > 0$. There exists $\epsilon > 0$, depending only on $k_1, k_2, r_{\text{cvx}}(B)$, and the sectional curvature of (B, g), such if X is a vector field X satisfying $\|(\nabla X)^{-1}\| < k_1$, $\|\nabla\nabla X\| \leq k_2$, and $\|X\| \leq \epsilon$ pointwise on B, then X has a unique zero in B, namely $\lim_{n\to\infty}(\Phi_X)^n(q)$ for any $q \in B$.

Remark 2.7 We intentionally avoid assuming that that (M, g) is complete or has positive injectivity radius. In the application to the set of smooth points of the shape space Σ_n^k, if $n \geq 3$ then (M, g) is a dense open subset of a non-smooth real algebraic variety (cf. [2]), hence neither complete nor of positive injectivity radius. However, any closed subset of M with positive injectivity radius will be complete. In particular this applies to the closures of all the balls considered in this paper.

[1] In the differential geometry literature there is little consistency in the meanings attached to the terms "convex set" and "strongly convex set". There is quite an array of criteria one can imagine demanding of a convex set; see Definition 3.1 for a few of these.

Corollary 2.6 follows immediately from the following more quantitative version.

Theorem 2.8 *Let $U \subset M$ be connected and let $|K| = |K|(U), \delta = \delta(U)$. Define the functions $\phi_-(\cdot)$ and $C_1(\cdot,\cdot)$ by (2.11) and (2.15). Assume either of the following sets of hypotheses:*

Case 1. *Y is a vector field defined on U and at each point of U we have $\|Y\| \le \epsilon_0 < r_{\rm inj}(U)$ and $\|\nabla Y + I\| \le \epsilon_1$. Define $\Psi_Y = \exp \circ Y$ as in (2.3).*

Case 2. *$k_1, k_2 > 0$, X is a vector field defined and uniformly nondegenerate on U, and at each point of U we have $\|(\nabla X)^{-1}\| \le k_1^{-1}$, $\|\nabla\nabla X\| \le k_2$, and $\|X\| \le \epsilon < k_1 r_{\rm inj}(U)$. Define $\Phi_X = \exp \circ(-(\nabla X)^{-1} \circ X)$ as in (1.1).*

Then:

(a) For all $p \in U$, in Case 1 we have

$$\|(\Psi_Y)_{*p}\| \le \kappa(\Psi_Y) := \phi_-(|K|^{1/2}\epsilon_0) + C_1(\delta, \epsilon_0)\epsilon_1, \tag{2.17}$$

while in Case 2

$$\|(\Phi_X)_{*p}\| \;\le\; \kappa(\Phi_X) := \phi_-(|K|^{1/2}\epsilon k_1^{-1}) + C_1(\delta, \epsilon k_1^{-1}) k_2 k_1^{-2}\epsilon. \tag{2.18}$$

In Case 1, let $F = \Psi_Y$; in Case 2 let $F = \Phi_X$. In Case 1 (respectively Case 2) if ϵ_0, ϵ_1 are small enough (resp., ϵ is small enough) that $\kappa(F) < 1$, then $F : (U, d_U) \to (M, d_M)$ is a contraction with constant $\kappa(F)$, and therefore has at most one fixed point in U. If U contains an open ball $B = B_\rho(p_0)$ on whose closure the distance functions d_U, d_M coincide (a condition satisied by every subset of $\overline{U}$ if U is convex), and if

$$\|Y(p_0)\| < (1 - \kappa(F|_B))\rho \qquad \textit{(in Case 1)}, \tag{2.19}$$

or

$$\|X(p_0)\| < (1 - \kappa(F|_B))k_1\rho \qquad \textit{(in Case 2)}, \tag{2.20}$$

then $F : U \to M$ has a unique fixed point, and this fixed point lies in B. Equivalently, the vector field Y in Case 1, or X in Case 2, has a unique zero in U, and this zero lies in B. Assuming (2.19) or (2.20) as appropriate, F preserves B, and the fixed point is $\lim_{n\to\infty} F^n(q)$ for any $q \in B$.

(b) If M is a locally symmetric space of non-negative curvature bounded above by Δ, then in (2.17) and (2.18), we can replace the right-hand sides by the smaller bounds

$$\kappa_{\rm sym+}(\Psi_Y) \;:=\; \phi_+(\Delta^{1/2}\epsilon_0) + \epsilon_1 \tag{2.21}$$

$$\text{and} \quad \kappa_{\rm sym+}(\Phi_X) \;:=\; \phi_+(\Delta^{1/2}\epsilon k_1^{-1}) + k_2 k_1^{-2}\epsilon \tag{2.22}$$

respectively, provided $\Delta^{1/2}\epsilon_0 < 3\pi/4$ in the first case and $\Delta^{1/2}\epsilon k_1^{-1} \le 3\pi/4$ in the second.

Proof: (a) Case 1. The bound (2.17) follows from Proposition 2.3, and (2.16). If $\kappa(\Psi_Y) < 1$ Lemma 2.2 implies that $\Psi_Y : (U, d_U) \to (M, d_M)$ is a contraction with constant κ. To use the fixed-point theorem we need a contraction with respect to a single distance function. However, the assumption that $d_U = d_M$ on B implies that the restriction of Ψ_Y to B is a κ-contraction from $(\overline{B}, d_M) \to (M, d_M)$. As noted in Remark 2.7, the metric space $(\overline{B}, d_M)$ is complete. Hence the result follows from Theorem 2.1 with $U = \overline{B}$.

Case 2. Letting $Y = -(\nabla X)^{-1}X$, for $p \in U$ we have $\|Y_p\| \leq \epsilon k_1^{-1} < r_{\text{inj}}(U)$, and from (2.5) we have $\|(\nabla Y + I)_p\| \leq k_1^{-2}k_2$. Hence Case 2 follows from Case 1.

(b) This follows from (2.12) and the proof of (a). ∎

Remark 2.9 In the bound (2.18), as either $\epsilon \to 0$ or $|K| \to 0$, we have $\phi_-(|K|^{1/2}\epsilon k_1^{-1}) \to 0$ and $C_1(\delta, \epsilon k_1^{-1}) \to 1$. Hence, as one would hope, for small ϵ and for small $|K|$ the bound (2.18) is asymptotic to $k_2 k_1^{-2}\epsilon$, the well-known bound for the Euclidean case (see the discussion following (2.7)).

Remark 2.10 Theorem 1.2 follows immediately from Case 2 of Theorem 2.8(a).

3 Averaging Points in a Riemannian Manifold

In its most elementary form, averaging is something that one does to a finite list of elements in a vector space. The average of a list $\{w_1, \ldots, w_m\}$ in a vector space V can be uniquely characterized as that vector $\overline{w} \in V$ for which

$$\sum_{i=1}^{m}(w_i - \overline{w}) = 0. \tag{3.1}$$

The "balancing property" (3.1) motivates the alternative term for the average, *center of mass*. If V is given *any* inner product then, using the the inner product to define a norm, the average above can also be uniquely characterized as

$$\overline{w} = \text{that vector } v \text{ which minimizes } \sum_{i=1}^{m} \|w_i - v\|^2$$

(the "least-squares property").

Unlike the balancing property, which requires a linear structure on V, the least-squares property makes sense if V is replaced by any metric space. A *Fréchet mean* of a finite subset of a metric space (A, d) is an element $a \in A$ at which the function $p \mapsto \sum_{q \in Q} d(q, p)^2$ attains an absolute minimum. In general a Fréchet mean need not

exist or be unique, but when it exists uniquely it is not unreasonable to call it the average of Q.

Modulo existence and uniqueness, Fréchet means give a way to extend the notion of "average" to finite lists of points in a Riemannian manifold, or more generally probability distributions on such a manifold. This idea of the *Riemannian center of mass* dates back at least as far as E. Cartan [3] in the case of simply connected manifolds of nonpositive curvature; in this setting the Fréchet mean of any probability distribution exists uniquely. However, the arbitrary-curvature case seems not to have been studied systematically until the 1970's, when it was investigated principally by Karcher and Grove ([16, 7, 12]; see also [14] §§4–5).

Unlike in Euclidean space, on a general Riemannian manifold it is clear that some restriction on the set of points to be averaged is necessary; for example there is no reasonable way uniquely to define the average of antipodal points on a sphere. Averaging can be done sensibly only on sets satisfying some suitable convexity condition (of which there are several). One notion of convexity was given in Definition 2.5; some other relevant notions are given below. The reader is warned that different authors attach different names to these notions.

Definition 3.1 Let $U \subset M$. We call U

- *self-visible* if any two points of U can be joined by at least one geodesic, not necessarily minimal, lying in U;
- *simple* if for any two points in U there is exactly one connecting geodesic lying in U;
- *solipsistically convex* if for any two points $p, q \in U$ there exists a connecting geodesic in U whose length is minimal among all connecting arcs lying in U (hence of length $d_U(p, q)$).

A function f defined on a self-visible set U is called *(strictly) convex on U* if its restriction to every geodesic in U is a (strictly) convex function of the arclength parameter.

If f is C^2 then a sufficient condition for f to be convex on U is that its covariant Hessian be positive-semidefinite on U; strict positivity implies strict convexity.

Definition 3.2 (cf. [14] p. 3) An open ball $B = B_\rho(p)$ is a *regular geodesic ball* if (i) $\rho < r_{\text{inj}}(p)$, and (ii)

$$\rho \cdot \max(0, \Delta(B))^{1/2} < \pi/2. \tag{3.2}$$

For $p \in M$ define the *regularity radius*

$$r_{\text{reg}}(p) := \sup\{\rho \mid B_\rho(p) \text{ is a regular geodesic ball}\}$$

and the *regular convexity radius*

$$r_{\text{regcvx}}(p) = \min(r_{\text{reg}}(p), r_{\text{cvx}}(p)).$$

For regular geodesic balls one has the following theorem of Jost [15]; see [14] Theorem 5.3 and [19] Theorem 1.7.

Theorem 3.3 *Let B be a regular geodesic ball in a complete Riemannian manifold. Then $\overline{B}$ is simple and solipsistically convex, and geodesics in B contain no pairs of conjugate points.*

Completeness of the ambient manifold is not essential in Theorem 3.3; if $B = B_\rho(p)$, it suffices that $\exp_p$ be defined on the closed ball of radius ρ centered at $0 \in T_pM$. The example of an open ball of radius π in the unit circle shows that a regular geodesic ball need not be convex. More generally Theorem 3.3 implies that regular geodesic ball $B \subset M$ is convex if and only if the distance functions d_B and d_M coincide on B.

There are various relations among $r_{\text{inj}}, r_{\text{cvx}}$, and r_{reg}; we mention only a few. By definition, $r_{\text{inj}}(p)$ is the largest of the three radii at p. If M is complete and has constant positive curvature, then Bonnet's Theorem ([4] Theorem 1.26(2)) implies that $r_{\text{cvx}}(p) \leq r_{\text{reg}}(p)$. But in general, a geodesic ball can be convex but not regular (see [11] for an example), or, as the circle example shows, regular but not convex.

Notation. If $p, q \in M$ can be joined by a unique minimal geodesic, we denote by $\exp_p^{-1}(q)$ the unique pre-image of q (under $\exp_p$) of smallest norm.

Now let Q be an arbitrary subset of a convex set $U \subset M$, and let μ be a probability measure on Q. For each $p \in U$ define

$$Y_Q(p) = \int_Q \exp_p^{-1}(q)\ d\mu(q)\ \in T_pM, \tag{3.3}$$

$$f_Q(p) = \frac{1}{2}\int_Q d(p,q)^2\ d\mu(q); \tag{3.4}$$

More properly these objects should be subscripted with the pair (Q, μ). However, in most of our results μ enters primarily through the geometry of Q rather than in the behavior of μ on Q. To emphasize this we will stick to the imperfect notation above.

Definition 3.4 Let $U \subset M$ be convex. (1) Let $Q \subset U$, let μ be a probability measure on Q, and define a vector field Y_Q by (3.3). If $Y_Q(p) = 0$ at a unique point $p \in U$, we call p the *(Riemannian) center of mass* of (Q, μ), *relative to* U. (2) Let $\tilde{Q} = \{q, \ldots, q_m\}$ be a finite list of points in U, let Q be the set of distinct elements of $\tilde{Q}$, let μ be the normalized counting-measure on Q, and define Y_Q as above. If $Y_Q(p) = 0$ at a unique point $p \in U$, we call p the *Riemannian average* of the list $\tilde{Q}$, *relative to* U.

We call a point a center of mass of (Q, μ) (respectively, a Riemannian average of the list $\tilde{Q}$) if it is the center of mass of (resp., Riemannian average) relative to some convex superset.

For a finite list $\tilde{Q}$, the definition of Riemannian average relative to U is simply the zero (assumed unique in U) of the vector field $Y = Y_{\tilde{Q}}$ on U defined by $Y(p) = \frac{1}{m}\sum_i \exp_p^{-1}(q_i) \in T_pM$. Since $\sum_i(\exp_p^{-1}(q_i) - Y_p) = 0$, heuristically, $Y(p)$ represents "balanced" average of the points q_i as seen from p. Alternatively, we can define $f_{\tilde{Q}} : U \to \mathbf{R}$, $f_{\tilde{Q}}(p) = \frac{1}{2m}\sum_{i=1}^m d(q_i,p)^2$, and assume that $f_{\tilde{Q}}$ is minimized uniquely at $\bar{q} \in U$. The Gauss Lemma ([4] p. 8]) implies that $\mathrm{grad}(d(q,\cdot))|_p = -\exp_p^{-1}(q)$, so $\mathrm{grad}(f_{\tilde{Q}}) = -Y_{\tilde{Q}}$, implying that $Y_{\tilde{Q}}$ has its zero at $\bar{q}$. Hence Definition 3.4 extends both the "balancing" and "least-squares" properties of the Euclidean average.

Remark 3.5 Definition 3.4 generalizes easily to a solipsistically convex or simple set U. In this case denote by $\exp_p^{-1,U}(q)$ that pre-image v of q (under $\exp_p$) of smallest norm for which $\exp_p(tv) \in U, 0 \le t \le 1$. In (3.3) we can replace $\exp_p^{-1}(q)$ by $\exp_p^{-1,U}(q)$, and $d(p,q)$ by $\|\exp_p^{-1,U}(q)\|$ (in the solipsistically convex case this is just $d_U(p,q)$). With these replacements it is still true that $\mathrm{grad}(f_Q) = -Y_Q$, but the interpretation of $Y_Q(p)$ as an average of points as seen from p is less compelling.

We will refine Definition 3.4 later for a case in which one center of mass is singled out, allowing us to dispense with the awkward "relative to U" (Definition 3.12).

Following [19, 23, 24], for example, we will call any relative minimum of f_Q a *Karcher mean*. Thus a Fréchet mean is necessarily a Karcher mean, but, absent extra hypotheses, not vice-versa. A center of mass of (Q,μ) under Definition 3.4 is simply a Karcher mean that lies inside some convex superset of Q.

Karcher proves a somewhat more general version of the following theorem ([16] Theorem 1.2, Definition 1.3, and Theorem 1.5).

Theorem 3.6 (Karcher) *Let (M,g) be a Riemannian manifold. Assume that $Q \subset B \subset M$, where $B = B_\rho(p_0)$ is a strongly convex ball. Let $\Delta = \Delta(B)$ be the supremum of the sectional curvatures in B. Then, with f_Q and Y_Q defined as above,*

(a) $\mathrm{grad}(f_Q) = -Y_Q$.

(b) The function f_Q achieves a minimum value on B, and hence Y_Q has a zero in B.

(c) If $\rho \cdot \max(0,\Delta(B))^{1/2} < \pi/4$, then the minimum of f_Q on B is achieved at a unique point $\overline{q}$, and for any point $p \in B$ we have

$$d(p,\overline{q}) \le \|Y_Q(p)\| \cdot \begin{cases} 1/h(\Delta, 2\rho) & \text{if } \Delta > 0 \\ 1 & \text{if } \Delta \le 0 \end{cases} \tag{3.5}$$

where, for $\Delta > 0$, $h(\Delta,x) = \Delta^{1/2}x\cot(\Delta^{1/2}x)$.

In [16], Karcher defines the center of mass to be the location of the minimum of f_Q on $\overline{B_\rho}$. However, his proof of existence and uniqueness of the minimum also implies uniqueness of the zero of Y_Q, so under the hypotheses of Theorem 3.6 this definition coincides with ours; indeed, the geometric Definition 3.4 is the one used in [7].

Note that the ball B in Theorem 3.6 is both geodesically convex and regular. If $\rho < \frac{1}{2} r_{\mathrm{reg}}(p_0)$ then the requirement on ρ in (c) is automatically satisfied; hence the upper limit on the radius of the balls for which part (c) is applicable is at least $\min(\frac{1}{2} r_{\mathrm{reg}}(p_0), r_{\mathrm{cvx}}(p_0))$ but no greater than $r_{\mathrm{regcvx}}(p_0)$. In [19] Theorem 7.3, W. S. Kendall strengthened the uniqueness assertion[2] in Theorem 3.6(c):

Theorem 3.7 (W. S. Kendall) *A mass distribution supported in a regular geodesic ball B has at most one Karcher mean in $\overline{B}$.*

In other words, as far as the uniqueness statement is concerned, as long as we assume $\rho < r_{\mathrm{inj}}(p_0)$ Karcher's $\pi/4$ can be replaced with $\pi/2$, and the ball $B_\rho(p_0)$ need not be assumed convex.

In general, Karcher means are not unique in the large, cf. [6, 20]. For example, given a set Q of two equally-weighted points in the unit circle S^1, the midpoints of each of the two arc joining the points is a Karcher mean. The statistically-natural absolute minimization of f_Q of course distinguishes one of these midpoints as the preferred one. However, we suggest an alternative, purely geometric way of distinguishing one of the Karcher means from the rest: just as in Euclidean space, the center of mass of a distribution μ should be in the convex hull, suitably defined, of its support—the average of a set Q should be not only near Q, but "within" Q. In the S^1 example above, unless the two points are antipodal—in which case the convex hull is not defined—only one of the two midpoints meets this criterion. Thus in this example the convex-hull and global-minimization criteria coincide, but the author does not know to what extent these criteria overlap in general.

The definition of "convex hull" varies in the literature. The notion best tailored to our needs is that of the *o-hull* defined below.

Definition 3.8 Call a set $Q \subset M$ *hulled* if it is contained in some convex set, and *o-hulled* if it is contained in some *open* strongly convex set. If Q is hulled (resp. o-hulled), define the *convex hull* of Q (respectively, the *convex o-hull* of Q), written $\mathrm{hull}(Q)$ (resp., $\mathrm{ohull}(Q)$) to be the intersection of all convex sets (resp. open strongly convex sets) containing Q. We will usually refer to these objects just as hulls and o-hulls.

Note that if a set is hulled, then the minimal geodesic between any two of its points exists and is unique.

Obviously hulls and o-hulls, when they exist, are convex sets, and $\mathrm{hull}(Q) \subset \mathrm{ohull}(Q)$. The o-hull may fail to exist even when the hull exists (example in S^1: a

[2]Kendall's proof does not yield existence of Karcher means as we have defined them. It is clear from the context and the proof that the existence asserted in the theorem as stated in [19] is the existence of a "solipsistic Karcher mean", in which the distance function $d = d_M$ in (3.4) is replaced by d_B. The existence argument requires $\mathrm{grad}(f_Q)$ to be outward-pointing on the boundary of the ball, which is guaranteed only under the solipsistic interpretation of f_Q.

semicircle closed at one endpoint and open at the other). However in $\mathbf{R}^n$, at least, the differences between hull and o-hull are minor: one always has

$$\text{hull}(Q) \subset \text{ohull}(Q) \subset \overline{\text{hull}(Q)} \tag{3.6}$$

(both inclusions can be strict; see [10]). Conceivably (3.6) holds generally for o-hulled sets in Riemannian manifolds provided hull(Q) has compact closure.

All sets Q of interest in this paper are contained in a convex open ball and so are o-hulled. As noted above, we will use o-hulls to distinguish one particular center of mass. Neither Karcher's theorem nor Kendall's generalization, as stated, immediately eliminates the unsettling possibility that Q could be contained in two different convex regular geodesic balls, and that Y_Q could have two zeroes (each of which could even be an absolute minimum of f_Q), each contained in one ball but not the other. However, the proofs in [16] and [19] imply more than is explicitly stated in either paper, and a minor extension of an ingredient of these proofs shows that this unwanted phenomenon cannot happen[3]. We give this extension in Lemma 3.10 and Corollary 3.11. The corollary leads us to the convex-hull criterion in Definition 3.12 below.

While ohull(Q) is the *smallest* set we can construct naturally from the family of open strongly convex supersets of Q, the *largest* set we can construct from this family also has relevance:

Definition 3.9 For any o-hulled set $Q \subset M$, define star(Q) to be the union of all open strongly convex supersets of Q. Analogously, define regstar(Q) to be the union of all regular geodesic balls containing Q. Note that star(Q) depends only on ohull(Q).

Given an open set $U \subset M$ and a boundary point $p \in \partial U$, call a tangent vector $v \in T_pM$ *outward-pointing* for U if $v \neq 0$ and if for some C^1 curve in M with $\gamma'(0) = -v$ we have $\gamma((0,\epsilon)) \subset U$ for some $\epsilon > 0$.

For reference, we record the following obvious facts (proof left to the reader).

Lemma 3.10 *Let $U \subset M$ be an open self-visible set with $\overline{U}$ compact, and let f be a C^1 function defined on some open neighborhood of $\overline{U}$.*

(a) If grad$f$ *is outward-pointing at each point of ∂U, then $f|_{\overline{U}}$ never achieves its minimum at a point of ∂U, and hence achieves it at some critical point $\overline{q} \in U$.*

(b) If f is convex on U then the critical points of f in U, if any, are global minima of $f|_{\overline{U}}$. If f is strictly convex on U then there is at most one critical point. ∎

[3] [23] uses a different partial solution to this problem: if in Karcher's theorem it is additionally assumed that $2\rho < r_{\text{inj}}(p_0)$ and hypothesis (c) is satisfied with $\Delta(B)$ replaced by $\Delta(B_{2\rho}(p_0))$—then Kendall's theorem implies that the Karcher mean of (Q,μ) in B_ρ is the unique Fréchet mean of (Q,μ). Our alternative approach does not require this extra hypothesis in order to single out a "best" Karcher mean, but our geometric definition of "best" differs from the statistical definition.

Corollary 3.11 *Let $Q \subset M$. Suppose that $f : M \to \mathbf{R}$ is C^1 on an open neighborhood of $\overline{\mathrm{star}(Q)}$ and that for every open strongly convex superset $U \supset Q$, the gradient of f is outward-pointing along ∂U. Suppose that there exists an open strongly convex superset $U_1 \supset Q$ of compact closure for which $f|_{U_1}$ achieves a minimum at some point $\overline{q}$, and that $\overline{q}$ is the unique local minimum of f in U_1. Then $\overline{q} \in \mathrm{ohull}(Q)$ and is the unique local minimum of f in $\mathrm{ohull}(Q)$. If $\mathcal{U}$ is any collection of supersets of Q on each of which f has a unique local minimum, then $\overline{q}$ is the unique local minimum of f in $\bigcup_{U \in \mathcal{U}} U$.*

Proof: Let $U \supset Q$ be open and strongly convex, with $\overline{U}$ compact. Then $U \bigcap U_1$ is an open strongly convex superset of Q, so ∇f is outward-pointing along $\partial(U \cap U_1)$, and $\overline{U \bigcap U_1}$ is compact. By Lemma 3.10, $f|_{U \cap U_1}$ achieves a minimum at some point q. But $\overline{q}$ is the unique local minimum of f in U_1; hence $q = \overline{q}$, so $\overline{q} \in U$ for every open convex superset of Q. ∎

In the case of our functions f_Q, the key point is that if U is an arbitrary open strongly convex superset of Q, then from (3.3) the vector field Y_Q is inward-pointing along ∂U, so $\mathrm{grad}(f_Q)$ is outward-pointing and Corollary 3.11 applies. Thus, while strongly convex or regular geodesic *balls* are essential to the proofs of Karcher's and Kendall's theorems (as well as to the proof of Theorem 4.8 in this paper), once one has existence and uniqueness within even one bounded strongly convex open ball, balls can essentially be dispensed with in favor of general strongly convex open sets. This allows us to frame our desired characterization of *the* center of mass, or *average*.

Definition 3.12 If (Q, μ) has a unique center of mass $\overline{q}$ in $\mathrm{ohull}(Q)$, we call $\overline{q}$ the *primary* center of mass, or simply *the* center of mass, of (Q, μ). If $\tilde{Q}$ is a finite list of points and μ is the normalized counting measure, we also refer to the primary center of mass as the *(Riemannian) average* of $\tilde{Q}$.

Thus, combining Theorems 3.6 and 3.7 with Corollary 3.11, we have the following.

Corollary 3.13 *Suppose $Q \subset M$ is contained in a strongly convex regular geodesic ball. Then for any probability distribution μ on Q, the primary center of mass $\overline{q}$ of (Q, μ) exists, lies in $\mathrm{ohull}(Q)$, and is the unique Karcher mean of (Q, μ) in $\mathrm{regstar}(Q)$. If f_Q has a local minimum at $\overline{q}$, then the restriction of f_Q to $\mathrm{regstar}(Q)$ achieves its absolute minimum at $\overline{q}$ and nowhere else.* ∎

In particular, Karcher means given by any two balls containing Q in Karcher's or Kendall's theorem coincide.

Note that $\mathrm{regstar}(Q)$ can be much larger than any single regular geodesic ball. For example, let M be the unit sphere S^n. Let $Q \subset S^n$ be a set of two non-antipodal points, let C be the minimal arc joining the points, and let C_{opp} be the arc antipodal to

C. Then $\mathrm{regstar}(Q) = \mathrm{star}(Q) = S^n - C_{\mathrm{opp}}$. In this and some other obvious examples on spheres, $\mathrm{regstar}(Q)$ coincides with $IC(Q)$:= the largest open superset of Q that does not meet the cut-locus of any point of $\mathrm{hull}(Q)$. It is plausible that in general $\mathrm{regstar}(Q) \subset IC(Q)$. However an example in [20] shows that in general $\mathrm{regstar}(Q)$ in Corollary 3.13 cannot be replaced by $IC(Q)$ in general without sacrificing uniqueness.

It is plausible that Corollary 3.13 remains true with "ohull" by "hull", but the author has not found a proof. However, Cheeger and Gromoll's general structure theorem for convex sets ([5] Theorem 1.6; note that our "convex" is Cheeger and Gromoll's "strongly convex") shows that $\mathrm{hull}(Q)$ has a well-defined dimension. Only if this dimension equals $\dim(M)$ is our definition of ohull exactly what is needed for the given proof of Corollary 3.13. However, Corollaries 3.11 and 3.13 can be sharpened to include the case $\dim(\mathrm{hull}(Q)) < \dim(M)$; see [10] (the original preprint version of this paper, available from the author).

4 Constructing the primary center of mass

The methods of §2 allow us to give a constructive proof of a version of Theorem 3.6. This section is devoted to the proof and a discussion of the consequences. Throughout we assume that the set Q lies in a strongly convex ball B.

The vector field Y_Q on B gives rise to a map $\Psi_Q = \Psi_{Y_Q} = \exp \circ Y_Q : B \to M$ as in Section 2. To apply our contracting-mapping result, Theorem 2.8, we need bounds on $\|\nabla Y_Q + I\|$. Heuristically it is easy to understand why this quantity is small, provided ρ is small enough. Let $\mathfrak{g}^{-1} : T^*M \otimes T^*M \to T^*M \otimes TM \cong \mathrm{End}(TM)$ be the isomorphism defined by using the metric to identify T^*M with TM ("raising an index" on the second factor of $T^*M \otimes T^*M$). For any function $f : M \to \mathbf{R}$, let $\mathrm{Hess}(f) = \nabla\nabla f \in \Gamma(\mathrm{Sym}^2 T^*M)$ denote its covariant Hessian, and let $\mathrm{Hess}'(f) = \mathfrak{g}^{-1}(\mathrm{Hess}(f)) \in \Gamma(\mathrm{End}(TM)$. From Theorem 3.6(a) we have $\nabla Y_Q = -\mathrm{Hess}'(f_Q)$. In normal coordinates $\{x^i\}$ centered at a point q, for points near q we have

$$\mathrm{Hess}(\frac{1}{2}r_q^2) = \sum_i dx^i \otimes dx^i \approx \sum_{i,j} g_{ij} dx^i \otimes dx^j = g,$$

so that $\mathrm{Hess}'(\frac{1}{2}r_q^2) \approx \mathfrak{g}^{-1}g = I$ near q. From [16] Theorem 1.5 we have

$$(\nabla Y_Q)(p) = -\int_Q \mathrm{Hess}'(\frac{1}{2}r_q^2)|_p \; d\mu(q). \tag{4.1}$$

Thus for general Q contained in a small set, at points near Q the endomorphism $-\nabla Y_Q$ is an average of endomorphisms close to the identity, and hence is close to the identity.

A quantitative bound on $\|\nabla Y_Q + I\|$ can be obtained in terms of the functions $h_\pm, h_0$ defined by

$$h_+(x) = x\cot x \ (0 \le x < \pi \text{ only}), \quad h_0(x) \equiv 1, \quad h_-(x) = x\coth x. \tag{4.2}$$

The function h_+ is monotone decreasing (hence ≤ 1), while h_- is monotone increasing (hence ≥ 1). Define

$$h(\lambda, r) = h_{\mathrm{sign}(\lambda)}(|\lambda|^{1/2} r) = \frac{\mathbf{c}(-\lambda r^2)}{\mathbf{s}(-\lambda r^2)}, \tag{4.3}$$

$$\psi(\lambda, r) = \mathrm{sign}(\lambda)(1 - h(\lambda, r)). \tag{4.4}$$

Then h is an analytic (entire) function of λr^2, with $h(\lambda, r) = 1 - \frac{1}{3}\lambda r^2 + O((\lambda r^2)^2)$. For every λ the function $r \mapsto \psi(\lambda, r)$ is nonnegative, monotone increasing on $[0, \pi)$ if $\lambda > 0$ and on $[0, \infty)$ if $\lambda \leq 0$, and $\psi(\lambda, r) = \frac{1}{3}|\lambda| r^2 + O(\lambda^2 r^4)$. For $\delta \leq \Delta \in \mathbf{R}$ and $0 \leq r < \pi \Delta^{-1/2}$ (the upper limit on r applying only if $\Delta > 0$), define

$$\psi_{\max}(\delta, \Delta, r) = \max(\psi(\Delta, r), \psi(\delta, r)) \tag{4.5}$$

$$= \frac{1}{3}|K| r^2 + O(|K|^2 r^4) \tag{4.6}$$

where $|K| = \max(|\delta|, |\Delta|)$. Note that $\psi_{\max}$ is monotone increasing in Δ and r, monotone decreasing in δ. Observing that $\frac{d^2}{dr^2}\psi(\pm 1, r) = \left\{ \begin{array}{c} 2\csc^2 r \\ 2\mathrm{csch}^2 r \end{array} \right\} \cdot \psi(\pm 1, r) \geq 0$, it also follows that $\psi_{\max}$ is a convex function of each argument with the other two held fixed. The relevance of $\psi_{\max}$ is in the following lemma.

Lemma 4.1 *Let $p, q \in M$ with $d(p, q) < r_{\mathrm{inj}}(q)$ and let δ and Δ be lower and upper bounds, respectively, for the sectional curvatures of M along the minimal geodesic from q to p γ; if $\Delta > 0$ also assume $d(p, q) < \pi \Delta^{-1/2}$. Then*

$$\|\mathrm{Hess}'(\frac{1}{2} r_q^2) - I\|(p) \leq \psi_{\max}(\delta, \Delta, d(q, p)). \tag{4.7}$$

If $d(p, q) \cdot \max(0, \Delta)^{1/2} < \pi/2$, then

$$\mathrm{Hess}(\frac{1}{2} r_q^2)|_p > 0. \tag{4.8}$$

Proof: Both statements follow immediately from Lemma 7.1 in the Appendix. ∎

Henceforth we assume that Q lies in a ball $B_D(p_0)$ and analyze the vector field Y_Q on a possibly larger concentric ball $B = B_\rho(p_0)$, still assumed strongly convex. We apply the lemma to points $p \in B, q \in Q$, setting $\delta = \delta(B), \Delta = \Delta(B)$. For such points we have $d(p, q) < \rho + D$, so to meet the potential restriction on $d(p, q)$ in the lemma, we assume that $(\rho + D) \max(0, \Delta)^{1/2} < \pi$. From (4.1), (4.7), and the monotonicity of $\psi_{\max}$ we then have

$$\|\nabla Y_Q + I\| = \| \int_Q (\mathrm{Hess}'(\frac{1}{2} r_q^2) - I)\, d\mu(q) \| \leq \psi_{\max}(\delta, \Delta, \rho + D). \tag{4.9}$$

We also have

$$\|Y_Q(p)\| \leq \sup_{q \in Q} d(p,q) \leq \rho + D. \tag{4.10}$$

Hence from Theorem 2.8, for all $p \in B$ we have

$$\|(\Psi_Q)_{*p}\| \leq \kappa(p_0; \rho, D) := \phi_\pm((\rho+D)|K|^{1/2}) + C_1(\delta, \rho+D)\ \psi_{\max}(\delta, \Delta, \rho+D). \tag{4.11}$$

where $|K| = |K|(B_\rho(p_0))$, and where the choice of sign in $\phi_\pm$ is governed by the following convention.

Notation Convention 4.2 *For the remainder of this paper, when an expression of the form $\phi_\pm(x)$ appears, $\phi_+(x)$ is to be used if M is a locally symmetric space of nonnegative curvature and $x \leq 3\pi/4$; $\phi_-(x)$ is to be used otherwise.*

To ensure that Ψ_Q is a contraction we want $\kappa(p_0; \rho, D) < 1$, which will be true for small ρ since $\phi_\pm(x)$ and $\psi_{\max}(\cdot, \cdot, x)$ are $O(x^2)$. This is not enough by itself to ensure existence of a fixed point:

Definition 4.3 Call a map $\Psi : (\mathrm{domain}(\Psi) \subset M) \to M$ *tethered to* Q if, for every strongly convex regular geodesic ball B containing Q, (i) Ψ is defined on B and (ii) $\Psi(B) \subset B$.

If we knew Ψ_Q to be tethered to Q (which implicitly requires $\mathrm{domain}(\Psi) \supset \mathrm{regstar}Q$), we could apply the general form of the Contracting Mapping Theorem (which assumes *a priori* that the contracting map preserves its domain) to conclude that Ψ_Q has a unique fixed point in $\overline{B_\rho(p_0)}$ as long as $\kappa(p_0; \rho, D) < 1$. In Euclidean space, Ψ_Q is always tethered to Q trivially: Ψ_Q maps the entire space to a single point contained in the convex hull of Q. On a general manifold, if Q consists of a single point then Ψ_Q is tethered to Q for the same trivial reason. Thus it seems likely that on general M, tethering will occur provided $\mathrm{diam}(Q)$ is sufficiently small. It is plausible that this happens for any Q contained in a strongly convex regular geodesic ball, but the author has neither a proof nor a counterexample. The lack of such a proof is the sole reason that in our center-of-mass application we use Theorem 2.1 (in the guise of Theorem 2.8) rather than the more general Contracting Mapping Theorem (but note that Theorem 2.8 may still be needed in other applications, i.e. those using maps Ψ_Y with Y not of the form Y_Q, since most such general maps will not be tethered). The cost is that the upper bound on the diameter of Q (or other measures of size such as the "circumradius") for which we can ensure that Ψ_Q has a fixed point is smaller than it would be if we knew that tethering occurred. Since it may be possible to prove tethering, either in general or in specific cases, in the remaining theorems of this paper we include statements of what one can conclude in the tethered case.

Assuming $\kappa(p_0;\rho,D) < 1$, to conclude from Theorem 2.8 that Ψ_Q has a fixed point, we additionally need to have

$$\|Y_Q(p_0)\| < (1-\kappa(p_0;\rho,D))\rho := s(p_0;\rho,D). \tag{4.12}$$

Clearly (4.10) is of no help here. However, the left-hand side of (4.12) does not depend intrinsically upon ρ, but only upon (Q,μ). We are taking $\rho \geq D$, so furthermore $s(p_0;\rho,D) \geq s(p_0;\rho,\rho) := s_2(p_0;\rho)$. The basis of the argument over the next few pages is simply that as long as $\|Y_Q(p_0)\|$ is less than the maximum value of the function $s_2(p_0;\cdot)$, there will be *some* radius ρ for which (4.12) is satisfied even with $D = \rho$, hence for all $D \leq \rho$ as well.

Note also that $\|Y_Q(p_0)\| \leq D$, so that an upper bound on D implies an upper bound on $\|Y_Q(p_0)\|$. Thus the most general conclusions we eventually draw will be those that have an upper bound only on $\|Y_Q(p_0)\|$ (hence on (Q,μ)) as a hypothesis, but as a corollary all such conclusions hold with an upper bound on D, a more easily checked and therefore more practical hypothesis. Eventually in Corollary 4.11 we will take p_0 to lie in Q, which will give us even more control since we can then take $D = \mathrm{diam}(Q)$.

Since we are interested not just in the *existence* of "good" radii ρ and D, but on estimating their size, we first prove a lemma establishing some properties of the function s; these will be used to estimate the size of balls on which Ψ_Q has a fixed point. In practice one is usually not presented with an explicit growth rate for $|\delta|, |\Delta|$, or $|K|$ as functions of ρ in (4.11), so we also examine the consequences of a (potentially less sharp but usually more practical version of the bound in (4.12), replacing the function s by a function $\tilde{s}$ defined below. The sharp bounds, however, are needed for the best estimates in [11] for an averaging algorithm on size-and-shape spaces.

Definition 4.4 Let $p \in M$. (a) For $0 \leq D \leq \rho < r_{\mathrm{reg}}(p)$, let $\Delta_{p,\rho} = \Delta(B_\rho(p))$, $\delta_{p,\rho} = \delta(B_\rho(p))$, $|K|_{p,\rho} = |K|(B_\rho(p))$, and

$$\kappa(p;\rho,D) = \phi_\pm((\rho+D)|K|_{p,\rho}^{1/2}) + C_1(\delta_{p,\rho},\rho+D)\psi_{\max}(\delta_{p,\rho},\Delta_{p,\rho},\rho+D), \tag{4.13}$$

$$s(p;\rho,D) = (1-\kappa(p;\rho,D))\rho. \tag{4.14}$$

(If $\delta_\rho = -\infty$ interpret (4.13) as $\kappa(p;\rho,D) = \infty$.)

(b) Let $r_1 \in (0, r_{\mathrm{reg}}(p))$, and let $\tilde{\Delta}(\cdot)$ (respectively $\tilde{\delta}(\cdot)$) be any continuous monotonically increasing (resp. decreasing) function on $[0,r_1]$ such that $\Delta_{p,\rho} \leq \tilde{\Delta}(\rho)$, $\delta_{p,\rho} \geq \tilde{\delta}(\rho)$, $r_1 \cdot \max(0,\tilde{\Delta}(r_1))^{1/2} < \pi/2$. For $0 \leq D \leq \rho \leq r_1$ define $\tilde{\kappa}(p,\tilde{\Delta},\tilde{\delta};\rho,D)$ to be the right-hand side of (4.13) with $\Delta_{p,\rho}, \delta_{p,\rho}, |K|_{p,\rho}$ replaced by $\tilde{\Delta}(\rho), \tilde{\delta}(\rho), \max(|\tilde{\Delta}(\rho)|, |\tilde{\delta}(\rho)|)$ respectively, and define

$$\tilde{s}(\rho,D) = \tilde{s}(\tilde{\Delta},\tilde{\delta};\rho,D) = (1-\tilde{\kappa}(\tilde{\Delta},\tilde{\delta};\rho,D))\rho. \tag{4.15}$$

In practice, $\tilde{\Delta}$ and $\tilde{\delta}$ will usually be *constant* functions, global upper and lower curvature bounds on $B_{r_1}(p)$. We define $\tilde{s}$ in greater generality above because this enables not only stronger results, but shorter proofs: anything proven for the more general functions $\tilde{s}$ applies to the special case $\tilde{s} = s$.

We construct from such a function $\tilde{s}$ several numbers and functions of D: $\tilde{D}_{\text{crit}}, \tilde{D}_{\max}$, and $\tilde{\rho}_i$, all defined below. The meaning of the $\tilde{\rho}_i(p, r_1; D)$ is indicated by the ρ_i in Figure 1; the qualitative correctness of Figure 1 is proven in Lemma 4.5.

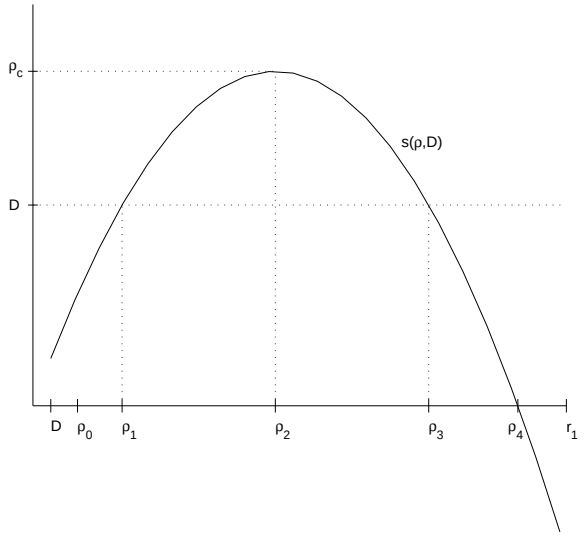

Figure 1. A (not-to-scale) sketch of $s(\rho, D)$ versus ρ for some fixed $D < D_{\text{crit}}$, assuming $r_1 > \rho_4$. A sketch of $\tilde{s}(\rho, D)$ for a fixed $D < \tilde{D}_{\text{crit}}$ would be similar, with the ρ_i replaced by $\tilde{\rho}_i, 1 \leq i \leq 4$. D_{crit} is the maximum value of $s(\rho, D_{\text{crit}})$; for $D > D_{\text{crit}}$, the graph of $s(\rho, D)$ lies entirely below the horizontal line at height D. To illustrate the maximum number of distinct radii we have sketched the case in which ρ_4 is strictly less than r_1, i.e. in which $\kappa(\rho, D)$ reaches 1 before ρ reaches r_1. The picture for smaller r_1 can be obtained from this one by moving r_1 to the left, say to $r_{1,\text{new}}$, truncating the diagram to the right of $r_{1,\text{new}}$ and decreasing D, if necessary, to keep it less than the maximum value of s on $[D, r_{1,\text{new}}]$ (hence keeping $r_{1,\text{new}} > \rho_1(D_{\text{new}})$). If any of ρ_2, ρ_3, ρ_4 in the picture above is to the right of $r_{1,\text{new}}$, the corresponding $\rho_{i,\text{new}}$ is defined to be $r_{1,\text{new}}$.

The definitions of $\tilde{D}_{\text{crit}}(p, r_1), \tilde{D}_{\max}(p, r_1)$ and the $\tilde{\rho}_i(p, r_1; \cdot)$ are given in (4.16–4.18) and (4.20–4.24) below. Here and below we suppress the parameters $\tilde{\Delta}$ and $\tilde{\delta}$ rather than write $\tilde{D}_{\text{crit}}(p, r_1, \tilde{\Delta}, \tilde{\delta})$ etc.; these parameters are always present implicitly. For the sharp-curvature-bound case ($\tilde{s} = s$), we omit the tildes and just write $D_{\text{crit}}(p, r_1), D_{\max}(p, r_1)$ and $\rho_i(p, r_1)$. Since $\tilde{\kappa}(\cdot, \cdot, 0, 0) = 0$, the sets over which the suprema are taken below are nonempty and the suprema well-defined.

$$\tilde{D}_{\max}(p, r_1) = \sup\{D \in [0, r_1] \mid \tilde{\kappa}(D, D) < 1\}. \tag{4.16}$$

$$\begin{aligned} D_{\max}(p) &= \sup\{D \in [0, r_{\text{reg}}(p)) \mid \tilde{\kappa}(D, D) < 1\} \\ &= \sup\{D_{\max}(p, r_1) \mid r_1 < r_{\text{reg}}(p)\}. \end{aligned} \tag{4.17}$$

$$\begin{aligned}
\tilde{D}_{\mathrm{crit}}(p, r_1) &= \sup\{D \in [0, r_1] \mid \exists \rho \in [D, r_1] \text{ for which } \tilde{s}(\rho, D) > D\}. && (4.18)\\
D_{\mathrm{crit}}(p) &= \sup\{D \in [0, r_{\mathrm{reg}}(p)) \mid \exists \rho \in [D, r_{\mathrm{reg}}(p)) \text{ for which } \tilde{s}(\rho, D) > D\} && (4.19)\\
&= \sup\{D_{\mathrm{crit}}(p, r_1) \mid r_1 < r_{\mathrm{reg}}(p)\}.
\end{aligned}$$

$$\tilde{\rho}_0(p, r_1; D) = \tilde{\rho}_0(p; D) = \begin{cases} D/h_+(2D\tilde{\Delta}(D)^{1/2}) & \text{if } \tilde{\Delta}(D) > 0, \\ D & \text{if } \tilde{\Delta}(D) \le 0. \end{cases} \qquad (4.20)$$

For $0 \le D < \tilde{D}_{\max}(p, r_1)$, define

$$\tilde{\rho}_4(p, r_1; D) = \sup\{\rho \in [D, r_1] \mid \tilde{\kappa}(\rho, D) < 1\}; \qquad (4.21)$$

for $0 \le D < D_{\max}(p)$ define

$$\rho_4(p; D) = \sup\{\rho \in [D, r_{\mathrm{reg}}(p) \mid \tilde{\kappa}(\rho, D) < 1\}. \qquad (4.22)$$

For $0 \le D < \tilde{D}_{\mathrm{crit}}$ define

$$\begin{aligned}
\tilde{\rho}_3(p, r_1; D) &= \sup\{\rho \in [0, r_1] \mid \tilde{s}(\rho, D) > D\}, && (4.23)\\
\tilde{\rho}_1(p, r_1; D) &= \inf\{\rho \in [0, r_1] \mid \tilde{s}(\rho, D) > D\}; && (4.24)
\end{aligned}$$

for $0 \le D < D_{\mathrm{crit}}$ define

$$\begin{aligned}
\rho_3(p; D) &= \sup\{\rho \in [0, r_{\mathrm{reg}}(p)) \mid \tilde{s}(\rho, D) > D\}, && (4.25)\\
\rho_1(p; D) &= \inf\{\rho \in [0, r_{\mathrm{reg}}(p)) \mid \tilde{s}(\rho, D) > D\}; && (4.26)
\end{aligned}$$

Note that for $i = 1, 3, 4$, $\rho_i(p; D)$ can alternatively be written as a supremum or infimum (over r_1) of $\rho_i(p, r_1; D)$ as we did above for $D_{\max}(p)$ and $D_{\mathrm{crit}}(p)$. Note also that in (4.16) and (4.21), "$\tilde{\kappa}(\cdot, \cdot) < 1$" can be replaced by "$\tilde{s}(\cdot, \cdot) > 0$" without altering the definitions of $\tilde{D}_{\max}$ and $\tilde{\rho}_4$.

The technical lemma below establishes some useful properties of the objects just defined, including monotonicity in parameters.

Lemma 4.5 *Let $p \in M$ and let $r_1 \in (0, r_{\mathrm{reg}}(p))$. Let $\tilde{\Delta}, \tilde{\delta}$ be continuous monotone bounds on curvature as in Definition 4.4(b), and let $\tilde{D}_{\mathrm{crit}} = \tilde{D}_{\mathrm{crit}}(p, r_1), \tilde{D}_{\max} = \tilde{D}_{\max}(p, r_1)$, and $\tilde{\rho}_i(\cdot) = \tilde{\rho}_i(p, r_1; \cdot)$ be as in (4.16–4.24).*

For $D \in [0, r_1]$ let $J_D = \{\rho \in [0, r_1] \mid D < \tilde{s}(\rho, D)\}$. For each D, the set J_D is either empty or an interval with endpoints $\tilde{\rho}_1(D), \tilde{\rho}_3(D)$. If $D_2 > D_1$ then $\overline{J_{D_2}} \subset J_{D_1}$, so $\{D \mid J_D \neq \emptyset\}$ is an interval whose right endpoint is $\tilde{D}_{\mathrm{crit}}$. $\tilde{D}_{\mathrm{crit}} > 0$ and $\bigcap_{0 \le D < \tilde{D}_{\mathrm{crit}}} J_D$ consists of a single point $\tilde{\rho}_{\mathrm{crit}}$, satisfying $\tilde{D}_{\mathrm{crit}} = \tilde{s}(\tilde{\rho}_{\mathrm{crit}}, \tilde{D}_{\mathrm{crit}}) = \max_{\rho \in [D, r_1]} s(\rho, \tilde{D}_{\mathrm{crit}})$, the maximum being achieved uniquely. The following are true:

1. $\tilde{D}_{\max} \geq \tilde{D}_{\mathrm{crit}}$, *with equality if and only if* $\tilde{D}_{\mathrm{crit}} = r_1$.

2. $D_{\mathrm{crit}} \geq \tilde{D}_{\mathrm{crit}}, D_{\max} \geq \tilde{D}_{\max}$.

3. $\rho_4(D) \geq \tilde{\rho}_4(D)$ *for all* $D < \tilde{D}_{\max}$.

4. $(2\tilde{D}_{\max}) \cdot \max(0, \tilde{\Delta}(\tilde{D}_{\max}))^{1/2} < \pi/2$.

5. *For each* $D \in [0, r_1]$, *the function* $\rho \mapsto \tilde{\kappa}(\rho, D)$ *on* $[D, r_1]$ *is continuous, monotone increasing, and convex. The function* $\rho \mapsto \tilde{\kappa}(\rho, \rho)$ *is* $O(|\tilde{K}|_\rho\ \rho^2)$, *where* $|\tilde{K}|_\rho = \max(|\tilde{\Delta}(\rho)|, |\tilde{\delta}(\rho)|)$.

6. *For each* $D \in [0, r_1]$, *the function* $\tilde{s}(\cdot, D)$ *is concave and achieves its maximum at a unique point* $\tilde{\rho}_2(D) \in (0, r_1]$.

For each $D < \tilde{D}_{\mathrm{crit}}$ *the following are true, where* $\tilde{\rho}_i = \tilde{\rho}_i(D)$.

7. $\rho_3 \geq \tilde{\rho}_3,\ \rho_0 \leq \tilde{\rho}_0$, *and* $\rho_1 \leq \tilde{\rho}_1$.

8. *The following order-relations hold (cf. Figure 1):*

$$D \leq \tilde{\rho}_0 \leq \tilde{\rho}_1 < \tilde{\rho}_{\mathrm{crit}} \leq \tilde{\rho}_3 < \tilde{\rho}_4 \leq \tilde{D}_{\max} \leq r_1. \tag{4.27}$$

9. $(\tilde{\rho}_4 + D) \cdot \max(0, \tilde{\Delta}(\tilde{\rho}_4))^{1/2} < \pi/2$.

As a special case, all conclusions above are true with the tildes erased. As a corollary, conclusion 4 is true also with $\tilde{D}_{\max}(p; r_1)$ *replaced by* $\tilde{D}_{\max}(p)$*; conclusions 5 and 6 are true with the tildes erased and with* $[D, r_1]$ *replaced by* $[D, r_{\mathrm{reg}}(p))$*; and conclusions 7–9 are true with* $\tilde{D}_{\mathrm{crit}}(p; r_1)$ *and* $\tilde{D}_{\max}(p, r_1)$ *replaced by* $D_{\mathrm{crit}}(p)$ *and* $D_{\max}(p)$ *respectively,* $\tilde{\rho}_i(p, r_1; D)$ *replaced by* $\rho_i(p; D)$*, and "*$\tilde{\rho}_4 \leq r_1$*" replaced by "*$\rho_4 < r_{\mathrm{reg}}(p)$*".*

Proof: From the definition of $\tilde{\kappa}$ continuity in all parameters is clear, and it is easy to check that $\tilde{\kappa}(\rho, D) \leq \tilde{\kappa}(\rho, \rho) = O(|\tilde{K}|_\rho\ \rho^2)$. We have already noted that $\psi_{\max}(\delta, \Delta, r)$ is monotone increasing in r and Δ, decreasing in δ, and convex in each variable separately; the same is true of $C_1(\delta, r)$. The functions $\phi_\pm$ are monotone increasing and convex. Monotonicity and convexity of $\phi_\pm, H$, and C_1 are retained after composition with the monotone functions $\tilde{\delta}(\rho), \tilde{\Delta}$.

It follows that with D held fixed, $\tilde{\kappa}(\cdot, D)$ is continuous, monotone increasing and convex, and hence that $\tilde{s}(\cdot, D)$ is continuous, concave, and, because of the factor of ρ in (4.15) and monotonicity, nonconstant on any interval of positive length. Since $\tilde{\kappa}(\rho, 0) = O(\rho^2)$, $\tilde{s}(\rho, 0) > 0$ for $\rho > 0$ sufficiently small. Hence J_0 is nonempty, and by continuity so is J_D for sufficiently small positive D. Hence $\tilde{D}_{\mathrm{crit}} > 0$.

For each fixed D, the concavity and local nonconstancy of the function $\tilde{s}(\cdot, D)$ implies that its maximum value $\tilde{\rho}_c(D)$ on $[0, r_1]$ is achieved at a unique point $\tilde{\rho}_2(D)$, and for any $a < \tilde{\rho}_c(D)$ the set $\{\rho \in [0, r_1] \mid \tilde{s}(\rho) > a\}$ is an interval; in particular each

set J_D is an interval. Since $D_2 > D_1$ implies $\tilde{s}(\rho, D_2) < \tilde{s}(\rho, D_1)$ strictly for $\rho > 0$, the asserted nesting of the intervals J_D also follows. The intersection of the nonempty J_D is nonempty because their closures are nested, and the intersection has only one point $\tilde{\rho}_{\rm crit}$ since $\tilde{s}(\cdot, D_{\rm crit})$ is nowhere constant. Continuity implies $\tilde{D}_{\rm crit} = \tilde{s}(\tilde{\rho}_{\rm crit}, \tilde{D}_{\rm crit})$.

From its definition clearly $\tilde{s}(\rho, D) \le \rho$. All the inequalities asserted in statement 8 follow immediately from the foregoing, except for $\tilde{\rho}_0 \le \tilde{\rho}_1$. The latter inequality follows from chasing through the definitions and monotonicity of the ingredients in $\tilde{\kappa}$. A helpful observation is that from (4.13) we have

$$\tilde{\kappa}(\rho, D) \ge \psi_{\max}(\tilde{\delta}(\rho), \tilde{\Delta}(\rho), \rho + D) \ge \psi(\tilde{\Delta}(\rho), \rho + D). \tag{4.28}$$

It also follows that $\tilde{s}(\tilde{D}_{\rm crit}, \tilde{D}_{\rm crit}) \ge \tilde{s}(\tilde{\rho}_{\rm crit}, \tilde{D}_{\rm crit}) = \tilde{D}_{\rm crit} > 0$, so that $\tilde{D}_{\max} \ge \tilde{D}_{\rm crit}$.

The monotonicity of $\phi_\pm, C_1$, and $\psi_{\max}$ imply that if $\rho \le r_1$, then $\kappa(\rho, D) \le \tilde{\kappa}(\rho, D)$, and hence $\tilde{s}(\rho, D) \le s(\rho, D)$. Hence $\tilde{D}_{\rm crit} \le D_{\rm crit}$, $\tilde{\rho}_1 \ge \rho_1$, and $\tilde{\rho}_i \le \rho_i$ for $i = 3, 4$.

To establish statements 4 and 9 we claim first that for $D < D_{\rm crit}$ we have

$$(\tilde{\rho}_1(D) + D) \max(0, \tilde{\Delta}(\tilde{\rho}_1(D)))^{1/2} < \pi/2. \tag{4.29}$$

This is true for $D = 0$, so if it is false for some $D < D_{\rm crit}$ then there exists $D \in (0, D_{\rm crit})$ for which $\tilde{\Delta}(\tilde{\rho}_1(D)) > 0$ and $(\tilde{\rho}_1(D) + D)\tilde{\Delta}(\tilde{\rho}_1(D))^{1/2} = \pi/2$, the latter implying $\psi(\tilde{\Delta}(\tilde{\rho}_1(D)), \tilde{\rho}_1(D) + D) = 1$. But the combination $D > 0, \tilde{\Delta} > 0$ implies strict inequality in (4.28), so $\tilde{\kappa}(\tilde{\rho}_1(D), D) > 1$ and $\tilde{s}(\tilde{\rho}_1(D), D) < 0$; but from the definition of $\tilde{\rho}_1$ we have $\tilde{s}(\tilde{\rho}_1(D), D) \ge D$. Hence (4.29) holds for all $D < D_{\rm crit}$. Therefore if statement 9 is false, there exists $\rho \in (\tilde{\rho}_1(D), \tilde{\rho}_4(D)$ for which $(\rho + D)\tilde{\Delta}(\rho)^{1/2} = \pi/2$. From (4.28) we again conclude that $\tilde{\kappa}(\rho, D) > 1$, and since $\rho \ge \rho_1(D) > 0$ this implies the strict inequality $\tilde{s}(\rho, D) < 0$, a contradiction since $\rho \in (0, \tilde{\rho}_4(D))$. This proves statement 9; a shorter version of the same argument yields statement 4. ∎

Remark 4.6 In Definition 4.4 and Lemma 4.5, the restriction "$r_1 < r_{\rm reg}(p)$" can be replaced by the less restrictive "$r_1 \cdot \max(0, \Delta_{p,r_1})^{1/2} < \pi/2$".

Corollary 4.7 *Let $p_0 \in M$, $0 < r_1 < r_{\rm regcvx}(p_0)$. Let $D_{\rm crit}, D_{\max}, \rho_4, \rho_1$ be as in (4.17)–(4.26). For $0 < \rho \le r_1$ write B_ρ for $B_\rho(p_0)$. Let $Q \subset B_{\rho_4}$ be equipped with a probability measure μ, and define Y_Q and f_Q by (3.3–3.4). Then Y_Q has at most one zero in B_{ρ_4} (equivalently, f_Q has at most one critical point in this ball); at such a zero f_Q achieves its minimum value on B_{ρ_4} (in fact, on* regstar(Q)*). If $D < D_{\rm crit}$ and $Q \subset \overline{B_D}$ (or more generally if $\|Y_Q(p_0)\| \le D$), then Y_Q has a unique zero $\overline{q}$ in B_{ρ_4}, and $\overline{q}$ lies in $\overline{B_{\rho_1}}$. Hence (Q, μ) has at most one center of mass in B_{ρ_4}, and has exactly one center of mass in B_{ρ_4} if $\overline{Q} \subset B_{D_{\rm crit}}$. If Ψ_Q is tethered to Q, these conclusions hold with $D_{\rm crit}$ replaced by the (never smaller and usually larger) number $D_{\max}$.*

We will prove this simultaneously with Theorem 4.8 below. But first, taking r_1 close to $r_{\rm regcvx}(p_0)$ in Corollary 4.7, note that Lemma 4.5 implies that the restriction on

the radius of the ball containing Q in Corollary 4.7 is more stringent than in Theorem 3.6(c). Similarly Lemma 4.5 implies that the conclusion $\overline{q} \in \overline{B_{\rho_1}}$ in the corollary above is not as sharp as Karcher's conclusion $\overline{q} \in \overline{B_{\rho_0}}$, and that the conclusion above concerning existence of at most one center of mass in B_{ρ_4} is weaker than Kendall's conclusion—at most one center of mass in $B_{r_{\rm reg}(p_0)}$—which is itself weaker than the uniqueness and minimization statement in Corollary 3.13. (However, we will see in §6 that if (M, g) has non-negative curvature, then for $D < \tilde{D}_{\rm crit}$ the uniqueness statement in Corollary 4.7 is actually stronger than Karcher's.) In fact, in view of Corollary 3.13, B_{ρ_4} can be replaced by regstar(Q) in the conclusions (but not the hypotheses) of Corollary 4.7.

Thus, were Corollary 4.7 the only outcome of the contracting-mapping approach, we would have gained little from it. However, the contracting-mapping approach additionally provides an *algorithmic construction* of the center of mass, one that is easily implemented in spaces for which the exponential map and its inverse are explicitly known, and in particular for shape spaces. In practice, any algorithm intended to average a list Q of points in a space is initialized at a point $q_0 \in Q$, but there are questions of whether the algorithm converges and whether its limit (if any) depends on the choice of initial point. As mentioned in the introduction, GPA algorithms converge quite rapidly in practical applications, but it is not readily apparent why this happens. For a given algorithm, one may be able to prove initial-point independence of the limit by one argument, and convergence by another, and perhaps estimate the convergence rate still another way. However, the contracting-mapping approach allows one to answer all these questions at once (although answering them individually by other means may lead to sharper answers, as in [23] Proposition 3, for initial-point independence in the GPA-S algorithm). Thus the added value of this approach lies in the following theorem, in which we state only those direct conclusions of the contracting-mapping approach neither contained in nor relying on Karcher's and Kendall's theorems (except for the use of ρ_0 in conclusion 3). In §5 we will see that by estimating the convergence rate of $\Psi_Q^n(p_0)$ and combining this with Kendall's uniqueness result, we can considerably strengthen certain parts of Theorem 4.8; see Theorem 5.3. In statement 3 of the theorem below, note that with the indicated restrictions on D, existence of the primary center of mass is guaranteed by Corollary 4.7, as well as by Theorem 3.6.

Theorem 4.8 *Let $p_0 \in M$, $0 < r_1 < r_{\rm regcvx}(p_0)$; for $0 < \rho \le r_1$ write B_ρ for $B_\rho(p_0)$. Let $\tilde{\Delta}(\cdot), \tilde{\delta}(\cdot)$ be continuous monotone upper and lower bounds on curvature as in Definition 4.4(b). Let $Q \subset B_{r_1}$ be equipped with a probability measure μ, and define Y_Q, f_Q by (3.3–3.4). Then, using the notation (4.16)–(4.26) with the parameter p_0 suppressed, the following are true.*

1. *$\tilde{D}_{\max}(r_1) \le D_{\max}(r_1) \le D_{\max}$ and $\tilde{D}_{\rm crit}(r_1) \le D_{\rm crit}(r_1) \le D_{\rm crit}$. In particular if $D < \tilde{D}_{\rm crit}(r_1)$ then all the $\rho_i(D)$ are defined, and*

$$D \le \tilde{\rho}_0(D) \le \tilde{\rho}_1(D) < \tilde{\rho}_{\rm crit} < \tilde{\rho}_3(D) \le \tilde{\rho}_4(D) \le \tilde{D}_{\max}(r_1) \le r_1 \qquad (4.30)$$

where $\tilde{\rho}_{\text{crit}}$ is value of ρ that maximizes $\tilde{s}(\rho, \tilde{D}_{\text{crit}})$.

2. *For all $D \in (0, r_1]$, if $Q \subset B_D$ and $\rho < \tilde{\rho}_4(D)$, then the map $\Psi_Q = \exp \circ Y_Q : B_\rho \to M$ is a contraction with constant $\tilde{\kappa}(p_0; \rho, D)$.*

3. *Assume that $Q \subset \overline{B_D}$ (or more generally that $\|Y_Q(p_0)\| \leq D$) and that either*

   (i) $D < \tilde{D}_{\text{crit}}$ and $\rho_1(D) < \rho < \rho_3(D)$, or

   (ii) $D < \tilde{D}_{\max}$, Ψ_Q is tethered to Q (Definition 4.3), and $D \leq \rho < \rho_4(D)$.

   Then Ψ_Q preserves preserves each ball B_ρ. In particular this holds for the D-independent radius $\tilde{\rho}_{\text{crit}}$. The sequence of iterates $\Psi_Q^n(q)$ converges to the primary center of mass $\overline{q}$ of (Q, μ) for every $q \in B_{\rho_3(D)}$ if (i) holds, and for every $q \in B_{\rho_4(D)}$ if (ii) holds. In either case $\overline{q}$ lies in $\overline{B_{\rho_0(D)}} \cap \text{ohull}(Q)$.

4. *For $D < \tilde{D}_{\text{crit}}$ the following relations hold:*

$$\rho_0(D) \leq \tilde{\rho}_0(D), \quad \rho_1(D) \leq \tilde{\rho}_1(D), \quad \rho_3(D) \geq \tilde{\rho}_3(D), \quad \rho_4(D) \geq \tilde{\rho}_4(D). \tag{4.31}$$

   If the curvature bounds $\tilde{\Delta}, \tilde{\delta}$ are taken to be constants (e.g. $\tilde{\Delta} \equiv \Delta(B_{r_1}), \tilde{\delta} \equiv \delta(B_{r_1})$), then the lower bound $\tilde{D}_{\text{crit}}$ on D_{crit} is a universal function of the numbers $r_1, \tilde{\Delta}$, and $\tilde{\delta}$, depending in no other way on the geometry of (M, g). Similarly the lower bounds $\tilde{D}_{\max}$ on $D_{\max}$, $\tilde{\rho}_i$ on ρ_i for $3 \leq i \leq 4$, and the upper bounds $\tilde{\rho}_i$ on ρ_i for $0 \leq i \leq 1$, are universal functions of $r_1, \tilde{\Delta}$, $\tilde{\delta}$, and D.

Remark 4.9 The chief point of the last two sentences in Statement 4 is that D_{crit}, the critical upper bound for D in Theorem 4.8, and $\rho_3(D)$, the radius of the ball on which the convergence in Statement 3 is guaranteed, are impossible to compute without knowing the functions $\rho \mapsto \delta(B_\rho), \rho \mapsto \Delta(B_\rho)$ precisely. Thus Statement 4 gives more easily used, if less sharp, lower bounds on these numbers. The analogous statement for ρ_1 will be used in §6 when we estimate the convergence rate of the sequence $\{\Psi_Q^n(p_0)\}$.

Remark 4.10 As $D \to 0$, the numbers $\rho_3(D)$ and $\rho_4(D)$ increase. Thus, the smaller the diameter of the set Q, the larger the set on which the theorem shows that the iterates Ψ_Q^n converge, and the larger the set on which the critical point of f_Q is guaranteed to be unique. Also note that $\lim_{D\to 0} \rho_3(D) = \lim_{D\to 0} \rho_4(D) = \sup\{\rho \in [0, r_1] \mid \kappa(\rho, 0) < 1\}$—a considerably larger number than $D_{\max} = \sup\{\rho \in [0, r_1] \mid \kappa(\rho, \rho) < 1\}$, which is the upper bound we would have found for the radii of the balls B_{ρ_3}, B_{ρ_4} in statement 3 and had we not separated the roles of the variables ρ and D in defining ρ_3 and ρ_4 (i.e. if we had used "2ρ" in place of "$\rho + D$" in (4.9) and (4.10)).

Proofs of Corollary 4.7 and Theorem 4.8: Statements 1 and 4 of the theorem just restate some of the conclusions of Lemma 4.5 for easy reference. Statement

2 follows from (4.11), since $s(\rho) > 0 \iff \kappa(\rho) < 1$. Statement 3 of Theorem 4.8 and the existence portion of Corollary 4.7 follow from Theorem 2.8 applied to $U = B_{\rho_4}, B = B_\rho$, since for $\rho_1 < \rho < \rho_3$ the fact that $D < s(\rho)$ ensures that the condition (2.19) is met. The conclusion that $\overline{q} \in \overline{B_{\rho_0}} \cap \mathrm{ohull}(Q)$ just combines Corollary 3.13 with Karcher's bound (3.5).

Integrating (4.8) over Q implies that $\mathrm{Hess}(f_Q) > 0$ on B_ρ provided that $(\rho + D) \cdot \max(0, \tilde{\Delta}(\rho))^{1/2} < \pi/2$, a condition that Lemma 4.5 (statement 9) ensures is met with $\rho = \tilde{\rho}_4$. Hence Lemma 3.10 implies that any critical point of f_Q in $B_{\tilde{\rho}_4}$ is unique and minimizes f_Q on this ball (in fact, on $\mathrm{regstar}(Q)$ by Corollary 3.13), proving the remainder of Corollary 4.7. ∎

Theorem 4.8 gives us an algorithm for computing the center of mass to any desired accuracy: start with some point q, and compute the iterates $\Psi_Q^n(q)$. As mentioned earlier, when Q is a finite set of points, it is natural to initialize the algorithm at some point of Q. This motivates the following corollary. In many cases of interest the ambient manifold is highly symmetric and the quantities $r_{\mathrm{regcvx}}(q), \tilde{D}_{\mathrm{crit}}(q)$ below are independent of q, enabling a much simpler statement of the corollary.

Corollary 4.11 *Let $Q \subset M$, μ a probability measure on Q. For simplicity let constants $\tilde{\Delta} \equiv \Delta(M), \tilde{\delta} \equiv \delta(M)$ be global upper and lower bounds on sectional curvature. For $q \in Q$ let $D_q(Q) = \sup\{d(q, q_1) \mid q_1 \in Q\}$, let $\tilde{D}_{\mathrm{crit}}(q) = \tilde{D}_{\mathrm{crit}}(q, r_{\mathrm{regcvx}}(q))$ be as in (4.18). If for at least one point $q_0 \in Q$ we have $D_{q_0}(Q) < \tilde{D}_{\mathrm{crit}}(q_0)$, then the center of mass $\overline{q}$ of (Q, μ) exists, and equals $\lim_{n\to\infty} \Psi_Q^n(q)$ for every $q \in Q$. In particular this conclusion holds for any $q_0 \in Q$ if $\mathrm{diam}(Q) < \tilde{D}_{\mathrm{crit}}(Q) := \inf\{\tilde{D}_{\mathrm{crit}}(q_0) \mid q_0 \in Q\}$.*

Proof: The hypotheses imply that $Q \subset B_D(q_0)$, where $D = D_{q_0}(Q)$. Letting $\epsilon = \tilde{D}_{\mathrm{crit}}(q_0) - D_{q_0}(Q)$ and defining $\tilde{\rho}_3 = \tilde{\rho}_3(q_0, r_{\mathrm{regcvx}}(q_0) - \epsilon/2; D)$ as in (4.23), we have $Q \subset B_{\tilde{\rho}_3}(q_0)$ since $D < \tilde{\rho}_3$. Hence statement 3 of Theorem 4.8 implies the result. ∎

Corollary 1.4 follows immediately.

Centering the underlying convex regular superdisk at a point of Q as in Corollary 4.11, while practical, is wasteful in terms of the restriction on the diameter of Q. Any set Q satisfying the hypotheses of Theorem 4.8 has a *(convex regular) circumradius* $\mathrm{circumrad}(Q)$: the supremum of the radii of open, strongly convex, regular geodesic balls containing Q. For $\mathrm{diam}(Q)$ sufficiently small (in particular, if Q admits a convex regular superdisk centered at one of its points) $\mathrm{circumrad}(Q) < \mathrm{diam}(Q)$, and the conclusion of Corollary 4.11 remains valid if $\mathrm{diam}(Q)$ is replaced by $\mathrm{circumrad}(Q)$ and if $\tilde{D}_{\mathrm{crit}}(Q)$ is replaced by $\tilde{D}_{\mathrm{crit}}(p_0)$, where p_0 is the "circumcenter". As a practical matter, the circumcenter is no easier to find than the center of mass, so that this strengthening of Corollary 4.11 is only useful if one has a uniform bound on $r_{\mathrm{regcvx}}(p)$ (and therefore on $\tilde{D}_{\mathrm{crit}}(p)$) for p in an appropriate neighborhood of Q. We will discuss this more quantitatively in §6.

5 Rapid convergence of the algorithms

Given an iterable map F, let $\mathrm{It}(F)$ denote the algorithm "iterate F". Under any contracting-mapping algorithm, the sequence of successive distances from one point to the next converges geometrically. However, it is well known that Newton's method does even better; each successive distance is bounded by a constant times the square of the preceding one. In this section we examine the convergence rates of algorithms of the form $\mathrm{It}(\Psi_Y)$ and $\mathrm{It}(\Phi_X)$ in general (where Ψ_Y and Φ_X are as in Theorem 2.8), and of the averaging algorithm $\mathrm{It}(\Psi_{Y_Q})$ of Theorem 4.8 and Corollary 4.11 in particular. We will see that while the convergence rate of $\mathrm{It}(\Psi_Y)$ for general Y is only geometric (although with a smaller ratio than $\kappa(\Psi_Y)$), the algorithms $\mathrm{It}(\Phi_X)$—more closely related to the flat-space Newton's method—have the same quadratic behavior as their flat-space cousins. The averaging algorithm falls somewhere in between: we obtain only geometric convergence, but with a very small ratio, provided that $\mathrm{diam}(Q)$ is small enough.

Throughout this section, notation will be as in Theorem 2.8. We denote the sequence of iterates $\{\Psi_Y^n(p_0)\}$ or $\{\Phi_X^n(p_0)\}$ by $\{p_n\}$. For any algorithm of the form $\mathrm{It}(\Psi_Y)$, the following proposition shows that the rate at which $d(p_n, p_{n+1}) \to 0$ is completely controlled by bounds on $\nabla Y + I$.

Proposition 5.1 *Let U be a convex set preserved by Ψ_Y, let $p_0 \in U$, and for $n > 0$ let $p_n = \Psi_Y^n(p_0)$. Then*

$$d(p_{n+1}, p_n) \leq (\sup_{p \in U} \|(\nabla Y + I)_p\|)\ d(p_n, p_{n-1}). \tag{5.1}$$

Proof: From the definition of Ψ_Y, we have

$$d(p_{n+1}, p_n) = \|Y_n\|. \tag{5.2}$$

To analyze how $\|Y_n\|$ changes when we increment n, fix n and let $\gamma : [0,1] \to M$ be the geodesic from p_n to p_{n+1} with initial velocity Y_n; thus $p_{n+1} = \gamma(1)$, $Y_n = Y_{\gamma(0)}$, and $Y_{n+1} = Y_{\gamma(1)}$. Let $\mathcal{P}_{\gamma(t)\to\gamma(0)}$ denote the operator of parallel transport along γ, with direction reversed, from $\gamma(t)$ back to $\gamma(0)$, let $A_p = (\nabla Y + I)|_p \in \mathrm{End}(T_pM)$, and let $\epsilon_1 = \sup_{p\in U} \|A_p\|$. Then

$$\begin{aligned} \frac{d}{dt}(\mathcal{P}_{\gamma(t)\to\gamma(0)}(Y_{\gamma(t)}) + tY_{\gamma(0)}) &= \mathcal{P}_{\gamma(t)\to\gamma(0)}(\nabla_{\gamma'(t)} Y) + \gamma'(0) \\ &= \mathcal{P}_{\gamma(t)\to\gamma(0)}(A_{\gamma(t)}(\gamma'(t))), \end{aligned}$$

since γ' is parallel along γ, and hence

$$\mathcal{P}_{\gamma(t)\to\gamma(0)}(Y_{\gamma(t)}) + (t-1)Y_{\gamma(0)} = \int_0^t \mathcal{P}_{\gamma(t_1)\to\gamma(0)}(A_{\gamma(t_1)}(\gamma'(t_1)))dt_1. \tag{5.3}$$

The integrand is bounded in norm by $\|A_{\gamma(t_1)}\|\|\gamma'(t_1)\| = \|A_{\gamma(t_1)}\|\|Y_n\|$. Hence

$$\|Y_{\gamma(t)}\| = \|\mathcal{P}_{\gamma(t)\to\gamma(0)}(Y_{\gamma(t)})\| \leq (1 - t + \int_0^t \|A_{\gamma(t_1)}\|\, dt_1)\, \|Y_n\| \tag{5.4}$$

$$\leq (1 - t + \epsilon_1 t)\|Y_n\|. \tag{5.5}$$

Inserting $t = 1$ we find $\|Y_{n+1}\| \leq \epsilon_1\|Y_n\|$, and hence

$$d(p_{n+1}, p_n) \leq \epsilon_1 d(p_n, p_{n-1}). \tag{5.6}$$

■

Thus in algorithms of the form $\mathrm{It}(\Psi_Y)$, successive distances decrease geometrically, but with ratio ϵ_1—a number smaller than the contraction constant $\kappa(\Psi_Y)$ in (2.17), and one whose only dependence on curvature is through Y itself.

To analyze the algorithms $\mathrm{It}(\Phi_X)$, proceed as above but with $Y = -(\nabla X)^{-1}X$; continue writing $A = \nabla Y + I$. In this case, for $p \in U$ and $v \in T_pM$, from (2.5) we have $A_p(v) = B_p(v)(Y_p)$, where $B_p(v) = -((\nabla X)^{-1} \circ (\nabla_v \nabla X))|_p$. Thus, pointwise we have

$$\|A\| \leq k_3\|Y\| \tag{5.7}$$

where $k_3 = k_1^{-1}k_2$. Inserting this bound into (5.4) with $t = 1$, and using (5.5) in the new integrand, we obtain $\|Y_{n+1}\| \leq \frac{1}{2}k_3(\epsilon_1 + 1)\|Y_n\|^2$ where now $\epsilon_1 = k_1^{-1}\epsilon$. Thus, with $k_4 = k_3(\epsilon_1 + 1)/2$, we have

$$d(p_{n+2}, p_{n+1}) \leq k_4 d(p_{n+1}, p_n)^2, \tag{5.8}$$

the same quadratic falloff as in flat-space Newton's method.

Note that the preceding analysis applies to any algorithm for which (5.7) holds, a condition intermediate between Case 1 and Case 2 of Theorem 2.8.

The convergence rates of $\mathrm{It}(\Psi_Y)$ and $\mathrm{It}(\Phi_X)$ can also be compared as follows. With the constants as named above, assume that for Ψ_Y that$\epsilon_1 < 1$, and for Φ_X that $k_4\epsilon_1 < 1$. Then for the algorithm $\mathrm{It}(\Psi_Y)$, we have

$$d(p_{n+1}, p_n) \leq d(p_1, p_0)\epsilon_1^n < \epsilon_1^{n+1}, \tag{5.9}$$

whereas for $\mathrm{It}(\Phi_X)$ we have

$$d(p_{n+1}, p_n) \leq k_4^{-1}(k_4 d(p_1, p_0))^{2^n} < k_4^{-1}(k_4\epsilon_1)^{2^n} \tag{5.10}$$

(if $k_4 = 0$, interpret (5.10) as $d(p_{n+1}, p_n) = 0$.)

In the proof of the Contracting Mapping Theorem (Theorem 2.1), to obtain convergence of the sequence $\{p_n = F^n(p_0)\}$, it suffices to know that (i) $d(p_n, p_{n+1}) \leq \kappa d(p_{n-1}, p_n)$ for all $n \geq 1$, and (ii) $d(p_0, p_1) < (1 - \kappa)\rho$. One does not need to know that F is a contraction on the whole ball B unless one wants to prove uniqueness of the fixed point and convergence of the sequence with other starting points. Thus the analysis above leads immediately to the following existence/convergence theorem to supplement Theorem 2.8.

Theorem 5.2 *Let $B = B_\rho(p_0) \subset M$ be a convex ball. Assume either of the sets of hypotheses listed as "Case 1" and "Case 2" in Theorem 2.8, with U replaced by the ball B. In Case 1, let $F = \Psi_Y$; in Case 2 let $F = \Phi_X$. Assume in addition the following:*

Case 1. *Assume $\|Y(p_0)\| < (1-\epsilon_1)\rho$.*

Case 2. *Let $k_4 = k_1^{-1}k_2(k_1^{-1}\epsilon + 1)/2$ and if $k_4 \neq 0$ assume that*

$$\sum_{n=0}^{\infty} k_4^{-1}(k_4 k_1^{-1}\|X(p_0)\|)^{2^n} < \rho.$$

Then in each case the sequence $\{F^n(p_0)\}$ lies in B and converges to a fixed point of F that lies in B.

The distances $d(p_{n+1}, p_n)$ in Case 1 have the exponential falloff given by (5.9), and in Case 2 have the super-exponential falloff given by (5.10). ∎

Theorem 5.2 is most useful when one knows ahead of time that there is at most one fixed point. This is exactly the case for averaging algorithm $\mathrm{It}(\Psi_{Y_Q})$ used in §4, since we do not need the contracting-mapping apparatus to prove uniqueness—given existence, we already know from Kendall's theorem that if Q is contained in regular geodesic ball B then $\Psi_Q := \Psi_{Y_Q}$ has at most one fixed point in B. This leads immediately to the following strengthening of certain portions of Theorem 4.8.

Theorem 5.3 *Let $p_0 \in M$, $0 < r_1 \leq r_{\mathrm{regcvx}}(p_0)$; for $0 < \rho \leq r_1$ write B_ρ for $B_\rho(p_0)$. Let $\tilde{\Delta}(\cdot), \tilde{\delta}(\cdot)$ be continuous monotone upper and lower bounds on curvature as in Definition 4.4(b). Define numbers $\tilde{D}'_{\mathrm{crit}}, \tilde{D}'_{\max}, \tilde{\rho}'_{\mathrm{crit}}$ and $\tilde{\rho}'_i$ analogously to the numbers defined in Lemma 4.5, but with $\tilde{s}$ replaced by the function*

$$\tilde{s}_{\mathrm{seq}}(\tilde{\Delta}, \tilde{\delta}; \rho, D) = (1 - \tilde{\kappa}_{\mathrm{seq}}(\tilde{\Delta}, \tilde{\delta}; \rho, D))\rho \tag{5.11}$$

where

$$\tilde{\kappa}_{\mathrm{seq}}(\tilde{\Delta}, \tilde{\delta}; \rho, D)) = \psi_{\max}(\tilde{\delta}(\rho), \tilde{\Delta}(\rho), \rho + D). \tag{5.12}$$

Then Statements 1 and 4 of Theorem 4.8 hold with $\tilde{D}_{\mathrm{crit}}, D_{\mathrm{crit}}, \tilde{\rho}_i$, and ρ_i replaced by $\tilde{D}'_{\mathrm{crit}}, D'_{\mathrm{crit}}, \tilde{\rho}'_i$, and ρ'_i respectively. Assume that $Q \subset \overline{B_D}$ (or more generally $\|Y_Q(p_0)\| \leq D$) and that either

(i) $D < \tilde{D}'_{\mathrm{crit}}$, or

(ii) $D < \tilde{D}'_{\max}$ and Ψ_Q is tethered to Q (see Definition 4.3).

Then the sequence of iterates $\{\Psi_Q^n(p_0)\}$ converges to the primary center of mass $\overline{q}$ of (Q, μ), and $\overline{q}$ lies in $\overline{q} \in \overline{B_{\rho_0(D)}} \cap \mathrm{ohull}(Q)$. The entire sequence lies in $\overline{B_{\rho'_1(D)}}$ (hence in the D-independent ball $B_{\rho'_{\mathrm{crit}}}$) if (i) holds, and in $B_{\rho'_4(D)}$ if (ii) holds. ∎

We have a corresponding strengthening of Corollary 4.11:

Corollary 5.4 *Corollary 4.11 remains true if the numbers $\tilde{D}_{\rm crit}(q)$ are replaced by the larger numbers $\tilde{D}'_{\rm crit}(q)$ defined in Theorem 5.3.* ∎

For the map $\Psi_Q = \Psi_{Y_Q}$ used in Theorem 5.3 and Corollary 5.4 we have a bound on the endomorphism A that, while not as strong as (5.7), is better than for the general Ψ_Y. From (4.9) and (4.6), if $Q \subset B_D(p_0)$ then on $B_\rho(p_0)$ we have

$$\|A\| \leq \psi_{\max}(\delta(B_\rho(p_0)), \Delta(B_\rho(p_0)), \rho + D) = \frac{1}{3}|K|(\rho+D)^2 + O(|K|^2(\rho+D)^4) \quad (5.13)$$

where $|K| = \max(\delta(B_\rho(p_0)), \Delta(B_\rho(p_0))$. Initialize the algorithm at a point $p_0 \in Q$ as in Corollary 4.11, let $D = \mathrm{diam}(Q)$, and assume that $D < D_{\rm crit}(Q)$ as in the corollary. From Theorem 5.2 Ψ_Q preserves the convex ball $\overline{B_{\rho_1}(p_0)}$, where $\rho_1(D)$ is the smallest positive number ρ satisfying $s(\rho, D) = D$, and hence when applying the bound (5.13) in the analysis of $\{\Psi_Q^n(p_0)\}$ it suffices to take $\rho = \rho_1(D)$. Since $s(\rho, D) = \rho(1 - O(|K|(\rho + D)^2))$, for D small we have $\rho_1(D) = D(1 + O(|K|D^2))$. Thus

$$\|A\| \leq \frac{4}{3}|K|\mathrm{diam}(Q)^2 + O(|K|^2\mathrm{diam}(Q)^4), \quad (5.14)$$

which we can use for ϵ_1 in (5.6) and (5.9). Thus for any $\epsilon_2 > 0$, if $|K| \cdot \mathrm{diam}(Q)^2$ is small enough we have

$$\begin{aligned} d(p_{n+1}, p_n) &\leq (\frac{4}{3} + \epsilon_2)|K|\mathrm{diam}(Q)^2 d(p_n, p_{n-1}) & (5.15)\\ &= k_5\ \mathrm{diam}(Q)^2 d(p_n, p_{n-1}), & (5.16)\end{aligned}$$

so in place of (5.9) we can write

$$\frac{d(p_{n+1}, p_n)}{\mathrm{diam}(Q)^n} \leq d(p_1, p_0)(k_5 \mathrm{diam}(Q))^n. \quad (5.17)$$

In other words, as $\mathrm{diam}(Q) \to 0$, the falloff rate of successive distances in the averaging algorithm is geometric even relative to $\mathrm{diam}(Q)$. The bound (5.10) shows that we would get even faster convergence to the center of mass if we iterated the map Φ_{Y_Q} instead of Ψ_{Y_Q}. However, as a practical tool Φ_{Y_Q} has the disadvantage that one must compute and invert ∇Y_Q, which may be difficult even if M has constant curvature, whereas for many more general spaces the algorithm $\mathrm{It}(\Psi_{Y_Q})$ is easily programmable.

Remark 5.5 Since $A = \nabla Y + I$, for $\mathrm{diam}(Q)$ small we can think of (5.13) as asserting that the vector field Y_Q is, in some sense, very nearly linear. From this point of view it is no surprise that the convergence of the algorithm is so rapid—what we are using is almost Newton's method for an almost linear function.

As $D \to 0$, the bound (5.15) can be improved by using the circumradius of Q instead of its diameter in this estimate (see the discussion after Corollary 4.11). In $\mathbf{R}^n$, one always has $\mathrm{circumrad}(Q) \leq \sqrt{\frac{n}{2(n+1)}}\,\mathrm{diam}(Q)$, with a regular n-simplex an extremal configuration. In a general Riemannian manifold, if we restrict attention to sets Q contained in a subset U on which the there are bounds on the curvature and a positive lower bound on the injectivity radius, then as $D \to 0$ the number $\sup\{\mathrm{circumrad}(Q)/\mathrm{diam}(Q) \mid Q \subset U,\ 0 < \mathrm{diam}(Q) \leq D\}$ tends to its Euclidean value. Thus we obtain an asymptotic bound $\epsilon_1 \sim \frac{2}{3}\frac{n}{n+1}\Delta D^2$, where $n = \dim(M)$.

6 Averaging in the case of non-negative curvature

When (M, g) has curvature of a fixed sign, the definitions of the critical radii in Theorem 5.3 and Corollary 5.4 simplify, since we can globally replace $\psi_{\max}(\delta_{p,\rho}, \Delta_{p,\rho}, \rho + D)$ in (5.12) by either either $\psi(\tilde{\Delta}(\rho), \rho + D)$ or $\psi(\tilde{\delta}(\rho), \rho + D)$. In this section we assume that the curvature is non-negative, which is true in all shape spaces and size-and-shape spaces.

The goal of this section is to estimate the critical radii appearing in Theorem 5.3 as well as the convergence rate of the averaging algorithm (not merely the asymptotics of this rate as $\mathrm{diam}(Q) \to 0$). To simplify the estimates further, we will assume a uniform upper bound $\tilde{\Delta} \equiv \Delta$ on sectional curvature in all the balls that appear in this section, and a uniform lower bound r_1 on the regular convexity radius of the center of any such ball. We assume $\Delta > 0$ strictly since the flat case is not very interesting, the algorithm converging at the first iteration.

Notation in this section will be for the most part as in §§4–5, but it is convenient to define rescaled variables $\bar{\rho} = \Delta^{1/2}\rho, \bar{D} = \Delta^{1/2}D$, and a rescaled function $\bar{s} = \Delta^{1/2}\tilde{s}$ of the rescaled variables (where in the definition of $\tilde{s}$ we take $\tilde{\delta} \equiv 0, \tilde{\Delta} \equiv \Delta$). We also write $\bar{\kappa}$ for $\tilde{\kappa}$ expressed in terms of the rescaled variables. We suppress all the parameters except D and ρ in most formulas below.

Fix $p_0 \in M$ and let $x = \bar{\rho} + \bar{D}$. Then

$$\bar{\kappa}(\bar{\rho}, \bar{D}) = \hat{\kappa}(x) := \psi(1, x) = 1 - x\cot x = \frac{1}{3}x^2 + O(x^4) \tag{6.1}$$

and

$$\bar{s}(\bar{\rho}, \bar{D}) = (1 - \bar{\kappa}(\bar{\rho}, \bar{D}))\bar{\rho}. \tag{6.2}$$

Since $\tilde{\Delta}, \tilde{\delta}$ are constant, $\bar{s}$ is differentiable, so the rescaled pair $(\bar{\rho}_{\mathrm{crit}}, \bar{D}_{\mathrm{crit}})$ from Lemma 4.5 can be characterized as the unique solution of the system of equations

$$\bar{s}(\bar{\rho}, \bar{D}) = \bar{D}, \tag{6.3}$$

$$\frac{\partial \bar{s}}{\partial \bar{\rho}}(\bar{\rho}, \bar{D}) = 0 \tag{6.4}$$

in $(0,\pi/2)\times(0,\pi/2)$, provided that $\bar{\rho}_{\text{crit}}$ as defined this way is less than $\Delta^{1/2}r_1$. For this system of equations, Maple's `fsolve` routine[4] yields $\bar{\rho}'_{\text{crit}} \approx .6816 \gtrsim .2169\pi, \bar{D}'_{\text{crit}} \approx .3952 \gtrsim .1258\pi$. Thus

$$\tilde{\rho}'_{\text{crit}} \geq \min(r_1, .2169\pi\Delta^{-1/2}), \quad \tilde{D}'_{\text{crit}} \geq \min(r_1, .1258\pi\Delta^{-1/2}). \tag{6.5}$$

From these numbers we also compute $\tilde{\rho}'_4(\tilde{D}'_{\text{crit}}) \approx \min(r_1, 1.1566\Delta^{-1/2}) \approx .3682\pi\Delta^{-1/2}$. Centering all balls below at p_0 and writing B_ρ for $B_\rho(p_0)$, we recall what the numbers just computed tell us: from Theorem 5.3, for any (Q,μ) with $\overline{Q}$ in the ball of radius $\tilde{D}'_{\text{crit}}$, and any p in the ball of radius $\tilde{\rho}'_{\text{crit}}$, the sequence $\{\Psi_Q^n(p)\}$ converges to the primary center of mass of (Q,μ). If Ψ_Q is tethered to Q, to conclude convergence we need only assume that $\overline{Q}$ and p lies in the balls of radius $\tilde{D}'_{\text{crit}}$ and $\tilde{\rho}'_4(\tilde{D}'_{\text{crit}})$ respectively.

If $Q \subset B_D$ then as $D \to 0$, the algorithm converges on larger and larger sets, the balls of radius $\tilde{\rho}'_3(D)$ (or $\tilde{\rho}'_4(D)$ in the tethered case). These radii approach $\tilde{\rho}'_3(0) = \tilde{\rho}'_4(0) = \min(r_1, (\pi/2)\Delta^{1/2})$. Thus as $D \to 0$ we get convergence on balls of radius arbitrarily close to (but smaller than) the largest radius for which Kendall's theorem (Theorem 3.7) guarantees uniqueness of the center of mass.

Remark 6.1 Corollary 4.7, the existence/uniqueness theorem given by the contracting-mapping approach, guarantees existence of the center of mass of a distribution supported in a ball of radius $\tilde{D}'_{\text{crit}}$; in Karcher's result, the $.1258\pi$ in (6.5) is replaced by the better $\pi/4$. To compare the uniqueness statement in Corollary 4.7 with those of Karcher and Kendall, we cannot use the radii above, coming from Theorem 5.3, but must go back to those in Theorem 4.8. This has the effect of replacing $\psi(1,x)$ in (6.1) by $\phi_-(x) + \psi(1,x) = \frac{2}{3}x^2 + O(x^4)$. In this case we analogously compute $\tilde{D}_{\text{crit}} \approx \min(r_1, .0904\pi\Delta^{-1/2})$ and $\tilde{\rho}_4(\tilde{D}_{\text{crit}}) \approx \min(r_1, .2777\pi\Delta^{-1/2})$[5]. Corollary 4.7 implies that if $\overline{Q}$ is contained in the ball of radius $\tilde{D}_{\text{crit}}$, then (Q,μ) has a unique center of mass in the ball of radius $\tilde{\rho}_4(\tilde{D}_{\text{crit}})$. Thus in the non-negative curvature case, for $D < \tilde{D}_{\text{crit}}$ the contracting-mapping approach, while giving not as strong a uniqueness statement as in Kendall's theorem, gives a slightly stronger statement than in Karcher's original theorem, which has only $\pi/4$ in place of our worst-case constant $.2777\pi$.

We next estimate the convergence rate of $\{p_n = \Psi_Q^n(p_0)\}$, assuming that $\overline{Q}$ lies in the ball of radius $\tilde{D}'_{\text{crit}}$. From Theorem 5.3 the sequence stays in the ball of radius $\rho'_{\text{crit}}(D)$, on which, letting $A = \nabla Y_Q + I$ and writing $x_{\text{crit}} = \bar{\rho}'_{\text{crit}}(D) + \bar{D}'_{\text{crit}}$, the bound (5.13) gives

$$\|A\| \leq \psi_{\max}(0,1,x_{\text{crit}}) = \hat{\kappa}(x_{\text{crit}}) \lesssim .4202 \tag{6.6}$$

[4]All numerical calculations in this section were done with Maple.

[5]These numbers increase slightly if (M,g) is further assumed to be locally symmetric, since instead of $\phi_-(x)$ we can then use the smaller quantity $\phi_+(x) = \phi_-(x) - \frac{1}{15}x^4 + O(x^6)$. In this case we can replace $.0904\pi$ by $.0932\pi$, and $.2777\pi$ by $.2991\pi$. The improvement is so marginal because $\phi_+(x)$ and $\phi_-(x)$ differ by only $\frac{1}{15}x^4 + O(x^6)$.

Hence we obtain the geometric convergence rate (5.6) with $\epsilon_1 = .4202$.

If we start with $p_0 \in Q$ and assume $D = \mathrm{diam}(Q) < \tilde{D}'_{\mathrm{crit}}$ as in Corollary 5.4, then as D decreases we can sharpen the convergence-rate estimate by replacing x_{crit} with $\bar{\rho}'_1(D) + \bar{D}$ in the previous estimate. Since $\hat{\kappa}$ is monotone increasing on $[0, x_{\mathrm{crit}}]$, and $\tilde{s}(\rho_1(D), D) = D$, we have $\rho_1(D) \leq \frac{D}{1-\hat{\kappa}(x_{\mathrm{crit}})} := c_1 D \leq 1.725D$. The function $x \mapsto \hat{\kappa}(x)/x^2$ is monotone increasing on $[0, \pi)$, so for $x \in [0, x_{\mathrm{crit}}]$ we have $0 \leq \hat{\kappa}(x) \leq (\hat{\kappa}(x_{\mathrm{crit}})/x^2_{\mathrm{crit}})x^2 := c_2 x^2$. Thus $\|A\| \leq c_2(1+c_1)^2 \Delta D^2 \leq 2.690 \Delta D^2$, so we can take $\epsilon_1 = 2.690 \Delta D^2$ in (5.6) and (5.9).

As $D \to 0$, this can be improved further—(5.15) gives a bound on ϵ_1 asymptotic to $\frac{4}{3}\Delta D^2$, and as noted at the end of §5 this can even be reduced to $\frac{2}{3}\frac{n}{n+1}\Delta D^2$, where $n = \dim(M)$.

Finally, we consider two simple examples: round spheres and complex projective spaces, with standard metrics. If M is a round sphere of radius R, then the curvature is constant and equal to R^{-2}, and $r_{\mathrm{cvx}}(M) = r_{\mathrm{reg}}(M) = \pi R/2$. Hence we can take $\Delta^{-1/2} = R$ and erase "min","r_1" and the tildes in all the estimates above; e.g. in place of (6.5) we have simply

$$\rho'_{\mathrm{crit}} \geq .2169\pi R, \quad D'_{\mathrm{crit}} \geq .1258\pi R \tag{6.7}$$

Similarly, $\mathbf{C}P^n$ with a Fubini-Study metric (unique up to scale) is a symmetric space of positive curvature. If we fix the scale by taking the metric to be the one for which the standard projection from the unit sphere $S^{2n+1} \to \mathbf{C}P^n$ is a Riemannian submersion, then the sectional curvatures of $\mathbf{C}P^n$ run between $\delta = 1$ and $\Delta = 4$ if $n \geq 2$ (the curvature is identically 4 if $n = 1$; $\mathbf{C}P^1$ with this metric is a round sphere of radius 1/2). In this case we have $r_{\mathrm{cvx}}(M) = \pi/4$ and $\Delta^{-1/2} = 1/2$, so the critical radii are exactly half those for the unit sphere; bounds are given by (6.7) with $R = 1/2$. It is not hard to show that Σ^k_2, the shape space of k points in $\mathbf{R}^2$, is exactly $\mathbf{C}P^{k-2}$ with this metric (if $k > 2$) [17], so the numbers above directly relate to the behavior of the Riemannian averaging algorithm on this shape space.

7 Appendix

7.1 Proof and discussion of Proposition 2.3

In this subsection, hypotheses and notation are as in Proposition 2.3. We first prove (2.10) and then discuss how to sharpen this bound for locally symmetric spaces; the bound (2.12) follows as a special case of this discussion.

Proof of (2.10). $J^{\|}$ and $J^{\perp}$, the components of $\hat{J}_v$ parallel and perpendicular to γ', are themselves Jacobi fields, with $J^{\|}(t) = (at + c)\gamma'(t)$ for some $a, c \in \mathbf{R}$. Each of $J^{\|}$ and $J^{\perp}$ satisfies antidiagonal initial conditions. In particular, $c = -a$, so $J^{\|}(1) = 0$. Hence $\hat{J}_v(1) = J^{\perp}(1)$, so it suffices to prove (2.10) under the assumption that $v \perp \gamma'(0)$, which we make henceforth.

Let $\{e_i\}_0^{n-1}$, where $n = \dim(M)$, be an orthonormal basis of T_pM with $e_0 = \gamma'(0)/\|\gamma'(0)\|$, and extend each e_i along γ by parallel translation. Write $J(t) = \sum_{i=1}^{n-1} f^i(t)e_i(t)$ and let $f : [0,1] \to \mathbf{R}^{n-1}$ be the vector-valued function whose components are the f^i; note that $\|f(t)\|_{\text{Euclidean}} = \|J(t)\|$. Then (2.2) simply becomes

$$f''(t) = A(t)f(t) \tag{7.1}$$

for a certain $(n-1)\times(n-1)$ matrix-valued function A whose operator norm satisfies $\|A(t)\| \le |K|(\gamma(t))\|\gamma'(t)\|^2$. The norm of $\gamma'(t)$ is constant and equal to the length r of γ. Letting $b = |K|(\gamma)$, we therefore have $\|A(t)\| \le br^2$.

For $v \in T_pM$ write $v = \sum v^i e_i$, and let $\bar{v} \in \mathbf{R}^n$ be the vector whose components in the standard basis are the v^i. The initial conditions for $\hat{J}_v$ then become $f(0) = -f'(0) = \bar{v}$. The unique solution of (7.1) with these initial conditions is given explicitly by the series

$$\begin{aligned} f(t) &= (1-t)\bar{v} + \int_0^t \int_0^{t_2} (1-t_1)A(t_1)\bar{v}\; dt_1 dt_2 \\ &\quad + \int_0^t \int_0^{t_4} \int_0^{t_3} \int_0^{t_2} (1-t_1)A(t_3)A(t_1)\bar{v}\; dt_1 dt_2 dt_3 dt_4 \\ &\quad + \ldots + \int \ldots \int_{0\le t_1 \le t_2 \ldots \le t_{2m} \le t} (1-t_1)A(t_{2m-1})A(t_{2m-3})\ldots A(t_1)\bar{v}\; dt_1 \ldots dt_{2m} \\ &\quad + \ldots \end{aligned} \tag{7.2}$$

(This series converges in norm uniformly on any compact t-interval.) In the $2m$-fold integral the integrand is bounded in norm by $(1-t_1)b^m r^{2m}\|v\|$ provided $0 \le t \le 1$, the only case we are interested in. Integrating explicitly, we obtain $\|v\|b^m r^{2m}(\frac{t^{2m}}{(2m)!} - \frac{t^{2m+1}}{(2m+1)!})$ as an upper bound on the $2m$-fold integral. Hence for $0 \le t \le 1$ we have

$$\begin{aligned} \|f(t)\| &\le \sum_{m=0}^{\infty} b^m r^{2m}\left(\frac{t^{2m}}{(2m)!} - \frac{t^{2m+1}}{(2m+1)!}\right)\|v\| \\ &= \left(\cosh(b^{1/2}rt) - \frac{\sinh(b^{1/2}rt)}{b^{1/2}r}\right)\|v\|. \end{aligned}$$

Plugging in $t = 1$, the bound (2.10) follows. ∎

In contrast to more frequently-seen bounds on Jacobi fields, the sign of the sectional curvature does not play a role in (2.10). The reason is the anti-diagonal initial condition, which in Euclidean space leads to $J(1) = 0$. If M is positively curved, then $\|J\|$ can reach 0 before time 1 and then grow again, so that $\|J(1)\|$ cannot be bounded by its Euclidean analog. However, while it is not obvious how to get the best bound

in Proposition 2.3 for general manifolds, or even for nonnegatively curved manifolds, the analysis simplifies considerably for locally symmmetric spaces (manifolds whose Riemann tensor is covariantly constant; examples are S^n and $\mathbf{C}P^n$). In this case the matrix $A(t)$ in (7.1) is a constant symmetric matrix $r^2\hat{A}$, and the solution (7.2) collapses to

$$f(t) = (\mathbf{c}(t^2r^2\hat{A}) - t\ \mathbf{s}(t^2r^2\hat{A}))\bar{v} \tag{7.3}$$

(see Table 1 in §2.) Hence in this case (2.10) can be improved to

$$\|\hat{J}_v(1)\| \le \|\mathbf{c}(r^2\hat{A}) - \mathbf{s}(r^2\hat{A})\|\ \|v^\perp\|. \tag{7.4}$$

We can always choose an orthonormal basis in which the matrix $\hat{A}$ in the proof above is diagonal, say $\hat{A} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n-1})$. Then $\mathbf{c}(r^2\hat{A}) - \mathbf{s}(r^2\hat{A})$ becomes a diagonal matrix with entries $\mathrm{sign}(\lambda_i) \cdot \phi_{\mathrm{sign}(\lambda_i)}(|\lambda_i|^{1/2}r)$. The sectional curvatures of M range between $\delta \le \min\{\lambda_i\}$ and $\Delta \ge \max\{\lambda_i\}$ (we would have equality here if we replaced δ and Δ by the minimum and maximum sectional curvatures achieved on 2-planes tangent to γ) and $\phi_\pm$ are increasing functions on appropriate intervals: ϕ_- on $[0, \infty)$ (the Taylor coefficients are all nonnegative), ϕ_+ on $[0, x_0]$, where $x_0 \approx 0.87\pi$ is the first positive solution of $(x^2 - 1)\sin x + x\cos x = 0$. Hence

$$\|\mathbf{c}(r^2\hat{A}) - \mathbf{s}(r^2\hat{A})\| \le \begin{cases} \phi_+(\Delta^{1/2}r) & \text{if } 0 \le \delta \le \Delta \text{ and } \Delta^{1/2}r \le x_0, \\ \max(\phi_-(|\delta|^{1/2}r), \phi_+(\Delta^{1/2}r)) & \text{if } \delta \le 0 < \Delta \text{ and } \Delta^{1/2}r \le x_0, \\ \phi_-(|\delta|^{1/2}r) & \text{if } \delta \le \Delta < 0. \end{cases} \tag{7.5}$$

Thus for a locally symmetric space we can replace $\phi_-(r|K|(\gamma)^{1/2})$ in (2.10) by the appropriate line of (7.5); the top line yields (2.12), since $x_0 > 3\pi/4$. (We chose $3\pi/4$ in Proposition 2.3 for simplicity. Values of ϕ_+ that equal or exceed 1 are irrelevant to us since in Theorem 2.8(b) they lead to a useless bound on κ. The first positive x for which $\phi_+(x) = 1$ is approximately $.74\pi$, so the restriction $\Delta^{1/2}r \le 3\pi/4$ more than suffices for our considerations.)

If M has *constant* curvature—i.e. all sectional curvatures are equal, say to Δ—then the matrix in (7.3) is a multiple of the identity, leading us to sharp equality. In this case $\hat{A} = -\Delta I$ so we obtain

$$\|\hat{J}_v(1)\| = \phi_\pm(|\Delta|^{1/2}r)\|v^\perp\| \tag{7.6}$$

where ϕ_+ is used if $\Delta \ge 0$, and ϕ_- if $\Delta < 0$.

7.2 The Hessian of the squared distance function

Good references for the material in this subsection are [14], §5 and [16], Appendix C.

The lemma below was used in Lemma 4.1 and Corollary 4.7. The useful bound (7.9) is essentially proven in [14] Chapters 4-5, but is not explicitly stated in this form. (Theorem 5.2 of [14] asserts an inequality that looks identical to (7.9), but because

Hildebrandt's goal in [14] is a simple upper bound that applies to *all* vectors, not just those orthogonal to γ', he imposes the requirement $\delta \leq 0$.) The block-diagonal decomposition of the Hessian indicated in the lemma must generally be used in order to get the sharpest estimates on $\|\nabla Y + I\|$ when Y is the gradient of a function of the form $p \mapsto \int_Q f(d(p,q))\, d\mu(q)$.

Lemma 7.1 *Let $p, q \in M$ with $d(p,q) < r_{\text{inj}}(q)$ and let $H = \text{Hess}(\frac{1}{2}r_q^2)|_p$. Let $\gamma : [0,1] \to M$ be the minimal geodesic from q to p, let u a unit vector tangent to γ at p, and let $V_p^\perp \subset T_pM$ be the orthogonal complement of* $\text{span}(u)$*. Let δ and Δ be lower and upper bounds, respectively, for the sectional curvatures of M along γ; if $\Delta > 0$ also assume $d(p,q) < \pi\Delta^{-1/2}$. Then for all $v \in V_p^\perp$ we have the following:*

$$\begin{aligned} H(u,u) &= 1, && (7.7)\\ H(u,v) &= 0, && (7.8)\\ h(\Delta, d(p,q))\|v\|^2 &\leq H(v,v) \leq h(\delta, d(p,q))\|v\|^2. && (7.9) \end{aligned}$$

Proof: Recall that for any function f, vectors $X, Y \in T_pM$, and an arbitrary smooth extensions of X, Y to vector fields on a neighborhood of p, the covariant Hessian H_f is given by

$$H_f(X,Z) = X(Z(f)) - (\nabla_X Z)(f). \tag{7.10}$$

Let $f = \frac{1}{2}r_q^2$, let X be an extension of the unit tangent vector field $\gamma'/\|\gamma'\|$ and let Z be an extension of $v \in V_p^\perp$ that is parallel along γ. Then (7.7) is trivial, and, since the Gauss lemma implies $Z(r_q) \equiv 0$ along γ, (7.8) is trivial as well. The bound (4.8) can derived from the normal-Jacobi-field estimate [14] Theorem 4.2, followed by rescaling the arclength parameter as at the bottom of [14] p. 53, and then restricting the proof of [14] Theorem 5.2 to the case of vectors orthogonal to the geodesic. ∎

References

[1] F. L. Bookstein: A hundred years of morphometrics. *Acta Zoologica Scientarium Hungaricae* **44** (1998), 7-59.

[2] T. K. Carne: The geometry of shape spaces. *Proc. London Math. Soc.* **16** (1989), 407-432.

[3] E. Cartan.: *Léçons sur la Géométrie des Espaces de Riemann.* Gauthier-Villars, Paris, 1928.

[4] J. Cheeger and D. G.Ebin: *Comparison Theorems in Riemannian Geometry.* North Holland/American Elsevier, Amsterdam, 1975.

[5] J. Cheeger and D. Gromoll: On the structure of complete manifolds of nonnegative curvature. *Ann. Math.* **96** (1972), 413-443.

[6] J. M. Corcuera and W. S. Kendall: Riemannian barycentres and geodesic convexity. *Math. Proc. Camb. Phil. Soc.* **127** (1999), 253-269.

[7] K. Grove and H. Karcher: Riemannian center of mass and mollifier smoothing. *Math. Zeit.* **132** (1973), 11-20.

[8] C. Goodall: Procrustes methods in the statistical analysis of shape. *J. R. Statist. Soc. B* **53** (1991), 285-339.

[9] J. C. Gower: Generalized procrustes analysis. *Psychometrika* **40** (1975), 33-51.

[10] D. Groisser: Newton's method, zeroes of vector fields, and the Riemannian center of mass. Preprint (2001).

[11] D. Groisser: On the convergence of some Procrustean averaging algorithms. Preprint (2001).

[12] K. Grove: Center of mass and G-local triviality of G-bundles. *Proc. Amer. Math. Soc.* **54** (1976), 352-354.

[13] S. Helgason: *Differential Geometry, Lie Groups, and Symmetric Spaces*. Academic Press, New York, 1978.

[14] S. Hildebrandt: Harmonic mappings of Riemannian manifolds. *Harmonic Mappings and Minimal Immersions* (ed. Giusti, E.), Lecture Notes in Mathematics 1161, Springer-Verlag, Berlin (1985), 1–117.

[15] J. Jost: Eine geometrische Bemerkung zu Sätzen über harmonische Abbildungen, die ein Dirichletproblem lösen. *Manuscripta Math.* **32** (1980), 51-57.

[16] H. Karcher: Riemannian center of mass and mollifier smoothing. *Commun. Pure and Appl. Math.* **30** (1977), 509–541.

[17] D. G. Kendall: Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.* **16** (1984), 81-121.

[18] D. G. Kendall and H. Le: The Riemannian structure of Euclidean shape spaces: a novel environment for statistics. *Ann. Stat.* **21** (1993), 1225-1271.

[19] W. S. Kendall: Probability, convexity, and harmonic maps. *Proc. London Math. Soc.* **61** (1990), 371-406.

[20] W. S.Kendall: The propeller: a counterexample to a conjectured criterion for the existence of certain convex functions. *J. London Math. Soc.* (2) **46** (1992), 364-374.

[21] J. T. Kent: The complex Bingham distribution and shape analysis. *J. Roy. Statist. Soc.* B **56** (1994), 285-299.

[22] J. T. Kent: New Directions in Shape Analysis. *The Art of Statistical Science*, ed. Mardia, K. V. John Wiley & Sons, 1992.

[23] H. Le: Mean size-and-shape and mean shapes: a geometric point of view. *Adv. Appl. Prob.* **27** (1995), 44-55.

[24] H. Le: On the consistency of Procrustean mean shapes. *Adv. Appl. Prob.* **30** (1998), 53-63.

[25] H. Le: Locating Fréchet means with application to shape spaces. *Adv. Appl. Prob.* **33** (2001), 324-338.

[26] L. Loomis and S. Sternberg: *Advanced Calculus*. Addison-Wesley, Reading, Massachusetts, 1968.

[27] P. D. Sampson, F. Bookstein, F. Sheehan, and E. Bolson: Eigenshape analysis of left ventricular outlines from contrast ventriculograms. In *Advances in Morphometrics*, L. F, Marcus et. al. ed., Plenum Press, New York (1996), 211-233.